Deep Multimodal Learning for Joint Textual and Visual Reasoning

Patrick Bordes

École Doctorale Informatique, Télécommunications et Électronique (Paris)

DOCTORAL THESIS

Deep multimodal learning for joint textual and visual reasoning

---

Patrick Bordes

A thesis submitted in fulfillment of the requirements
for the degree of Doctor of Computer Science

Tentative defense date: September 30th, 2020

Jury composed of:

Mr. Yannis Avrithis	INRIA Rennes	Reporter
Mr. Loïc Barrault	University of Sheffield	Reporter
Mrs. Diane Bouchacourt	Facebook	Examiner
Mr. Patrick Gallinari	Sorbonne Univ. — Criteo	Supervisor
Mrs. Catherine Pelachaud	ISIR	Examiner
Mr. Benjamin Piwowarski	Sorbonne University	Co-supervisor

ABSTRACT

In the last decade, the evolution of Deep Learning techniques to learn meaningful data representations for text and images, combined with an important increase of *multimodal* data, mainly from social network and e-commerce websites, has triggered a growing interest in the research community about the joint understanding of language and vision. The challenge at the heart of Multimodal Machine Learning is the intrinsic difference in semantics between language and vision: while vision faithfully represents reality and conveys low-level semantics, language is a human construction carrying high-level reasoning. Two categories of work can be distinguished in Multimodal Machine Learning: the first intends to solve multimodal tasks, such as Image Captioning; the second, which is the purpose of this thesis, leverages visual information to solve a textual task (and vice-versa).

One the one hand, language can enhance the performance of vision models. The underlying hypothesis is that textual representations contain visual information. We apply this principle to two Zero-Shot Learning tasks. In the first contribution on ZSL, we extend a common assumption in ZSL, which states that textual representations encode information about the visual appearance of objects, by showing that they also encode information about their visual surroundings and their real-world frequence. In a second contribution, we consider the *transductive* setting in ZSL, in which unknown classes and their corresponding images are known during training (but not their correspondence). We propose a solution to the limitations of current transductive approaches, that assume that the visual space is well-clustered, which does not hold true when the number of unknown classes is high. To do so, we use the CycleGAN model to align textual and visual distributions in an unsupervised fashion.

On the other hand, vision can expand the capacities of language models. We demonstrate it by tackling Visual Question Generation (VQG), which extends the standard Question Generation task by using an image as complementary input, by using visual representations derived from Computer Vision. We show that image parts can be represented as word representations in a neural text model (here, Transformers). In another contribution, we leverage visual information to enhance textual representations. We expand traditional approaches that only consider word embeddings, and show that sentence representations can also benefit from visual semantics.

Finally, we present research perspectives on Multimodal Machine Learning.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

AI	Artificial Intelligence
BoW	Bag-of-Words
BoVW	Bag of Visual Words
BLEU	BiLingual Evaluation Understudy
CBOW	Continuous Bag-of-Words
CCA	Canonical Correlation Analysis
CNN	Convolutional Neural Network
CLEVR	Compositional Language and Elementary Visual Reasoning diagnostics
CMPlaces	Cross-Modal Places
CR	Customer Reviews
CV	Computer Vision
DSM	Distributional Semantic Model
FR	First Relevant
GAN	Generative Adversarial Network
GPU	Graphics Processing Unit
GRU	Gated Recurrent Unit
HOG	Histogram of Oriented Gradient
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
LSTM	Long-Short Term Memory
METEOR	Metric for Evaluation of Translation with Explicit ORdering
MFR	Mean First Relevant
MSRP	Microsoft Research Paraphrase
MR	Movie Review
MRR	Mean Reciprocal Rank
MLP	Multi-Layer Perceptron
ML	Machine Learning
MPQA	Multi-Perspective Question Answering
NER	Named Entity Recognition

NLP	Natural Language Processing
METEOR	Metric for Evaluation of Translation with Explicit ORdering
PCA	Principal Component Analysis
POS	Parts-of-Speech
CIDEr	Consensus-based Image Description Evaluation
RNN	Recurrent Neural Network
RPN	Region Proposal Network
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
NLM	Neural Language Model
SUBJ	Subjectivity/Objectivity
SICK	Sentences Involving Compositional Knowledge
SIFT	Scale-Invariant Feature Transform
SNLI	Stanford Natural Language Inference
SST	Stanford Sentiment Treebank
STS	Semantic Textual Similarity
SVD	Singular Value Decomposition
SVM	Support Vector Machine
t-SNE	t-distributed Stochastic Neighbor Embedding
TF-IDF	Term Frequency-Inverse Document Frequency
VRD	Visual Relationship Detection
VQA	Visual Question Answering
ZSL	Zero-Shot Learning
T-ZSL	Transductive Zero-Shot Learning
VQG	Visual Question Generation
CMR	Cross-Modal Retrieval
VGL	Visual Grounding of Language
MMT	Multimodal Machine Translation
VAE	Variational Auto-Encoder

CHAPTER 1

INTRODUCTION

Contents

1.1 Context

Over the last decades, the quantitative explosion of numerical data has induced a rising interest for the fields of Artificial Intelligence (AI) and Machine Learning (ML). For example, it is estimated that about 2.5 trillion bytes are created, every day, on Earth. This data can take various forms: text messages in natural language (650 million tweets per day on Twitter), images (a billion per day on Facebook), videos (300 hours of uploaded videos on Youtube per hour), but also audio recordings, transactions, GPS signals, climatic measures, etc. This exponential evolution of generated data has gone hand in hand with an exponential evolution of computing power (Moore's Law), estimated at a 1 trillion-fold increase from 1956 to 2015. For example, a recent smartphone, with its integrated Graphics Processing Unit (GPU), has more computing power that the computer that made moon landing possible in 1969. Large amount of training data, combined with powerful computers, led to the development of Deep Learning (LeCun et al. 2015), a branch of Machine Learning focused on learning data representations (*representation learning*) with multiple layers corresponding to multiple levels of abstraction. Deep Learning techniques have been successfully applied to the two most common modalities: *text* and *image*[1].

Natural Language Processing (NLP) is a research field studying models and techniques able to process automatically textual data. NLP covers a variety of tasks that are based on textual data, such as sentiment analysis (Pang et al. 2007), Machine Translation (Bahdanau et al. 2014), summarization (Rush et al. 2015), etc. To tackle such tasks, automatic methods are required to capture textual syntax, semantics and lexical properties. The first methods, in the 1970s, relied

1. Throughout this thesis, we will only consider these two modalities.

on hand-coded set of rules for language (Winograd 1971): these rules were extremely costly to generate, error-prone, and generalized poorly. After that, statistical approaches modeled language as a probabilistic model with Markovian assumptions, where co-occurrence probabilities for words are estimated using counting techniques (Y. Bengio et al. 2003). However, such methods face important limitations, especially for rare words that may not be seen in the reference corpus. Instead, recent approaches based on Deep Learning have focused on learning textual *representations*: vectors that encode the semantics of linguistic units at several levels of granularity — e.g, word (Mikolov et al. 2013b), sentence (Kiros et al. 2015), document (Le et al. 2014) — that can be used on downstream NLP tasks. With representation learning, the semantic proximity between textual units is linked to a geometrical relationship in a vector space. Thus, building meaningful textual representations is at the core of NLP, as it directly influences performances on text-related tasks.

Understanding visual data from images is the purpose of Computer Vision (CV). CV covers a variety of tasks, such as classification (Deng et al. 2009), segmentation (Hariharan et al. 2014), or object detection (Sermanet et al. 2014). Traditional methods relied on hand-crafted features (Lowe 2004; Dalal et al. 2005), that are features obtained using an automatic method, without a learning phase. Such features were then incorporated in a standard statistical model, such as Support Vector Machine (SVM) (Tong et al. 2001). Here again, Deep Learning has enabled to build task-agnostic image representations, that successfully generalize to downstream tasks, with the use of Convolutional Neural Network (CNN). The recent success of CNNs in the last decade can be attributed to: new large-scale datasets such as ImageNet (Deng et al. 2009), technical improvements such as back-propagation (LeCun et al. 1989) and the rise in computing power. With CNNs, visual representations are *learned* from raw pixel data in a hierarchical way. While first layers detect low-level features like edges or corners, following layers progressively capture more high-level information, with more detailed elements (e.g, faces or animals). Similarly to NLP, the quality of visual representations is paramount in CV as they are used in vision-related downstream tasks.

While there is an abundant literature on textual and visual representation learning, the study of multimodal representation learning and multimodal tasks (i.e. tasks that require a joint understanding of text and vision) remain comparatively under-tackled. However, multimodal tasks have gained in importance in the last decade, with the emergence of large platforms with multimodal content, such as social networks (e.g, Facebook, Twitter, Instagram) or e-commerce websites (e.g, Amazon, Zalando). First concerns consisted in finding *alignment* between modalities, to perform Cross-Modal Retrieval: given an image, retrieve the text with the maximal semantic similarity, and vice versa. Now, a wide variety of complex multimodal tasks exist, that necessitate joint textual and visual reasoning, such as Image Captioning, where the goal is to generate a sentence, in natural lan-

guage, describing an image. In the present thesis, we will explore the current state of Multimodal Machine Learning, propose contributions on several multimodal tasks, and propose answers to central issues of Multimodal Machine Learning.

1.2 Research Questions

"What can be shown, cannot be said", writes Ludwig Wittgenstein in his *Tractatus logico-philosophicus* (Wittgenstein 1922). To *show* something, means referring to a *visual* scene, from the real-world experience. To *say* something about this scene, means using *language* to formulate an *abstract* description of this *concrete* situation. However, language will never be able to convey and describe all subtleties of the visual world. Indeed, language is a human construction, expressing *high-level* semantics, made by humans to address other humans. On the other hand, in comparison to text, vision faithfully depicts reality, is not subject to interpretation, and carries *low-level* semantics.

Linguistic and visual modalities present inherent differences (Grice 1975; Ahn et al. 2005). In an image, as processed by a computer, the constituting elements are the *pixels*: each pixel is defined by its color, given by RGB values, and corresponds to a light signal captured by a camera. Despite the fact that images possess a finite number of pixels, the information within a pixel is by nature *continuous*, since the content of a pixel could itself be sub-divided into a larger number of pixels, depending on the resolution of the camera that took the image. In contrast, the constituting elements of text are *words*, which are *discrete* by nature. Indeed, words are finite in a given language; for example, there is an estimated total of one million words in the English language, with about 170,000 words in current use, and about 30,000 words used by each individual.

The structure of both modalities is also intrinsically different. In images, pixels are understood *spatially*, in relation with the neighboring pixels in the four directions (up, down, right, left). At a higher granularity, the distance between objects in the image, which are themselves constituted of pixels, is directly linked to the real-world distance between them, with a factor depending on the orientation of the camera. On the other hand, language is by nature *sequential* and follows a unique direction, e.g., a sentence is read from right-to-left in English, and left-to-right in Arab. The distance between words in the sentence has no relation with the real-world, but is rather linked to the grammatical relation between words, e.g., pronouns tend to come before adjectives, that tend to come before nouns.

Due to the fact that language is a human construction, the information present in texts and images is intrinsically different. Indeed, there is a bias in language compared to vision, which is referred in the literature as the Human Reporting Bias (Gordon et al. 2013): *the frequency at which objects, relations, or events occur*

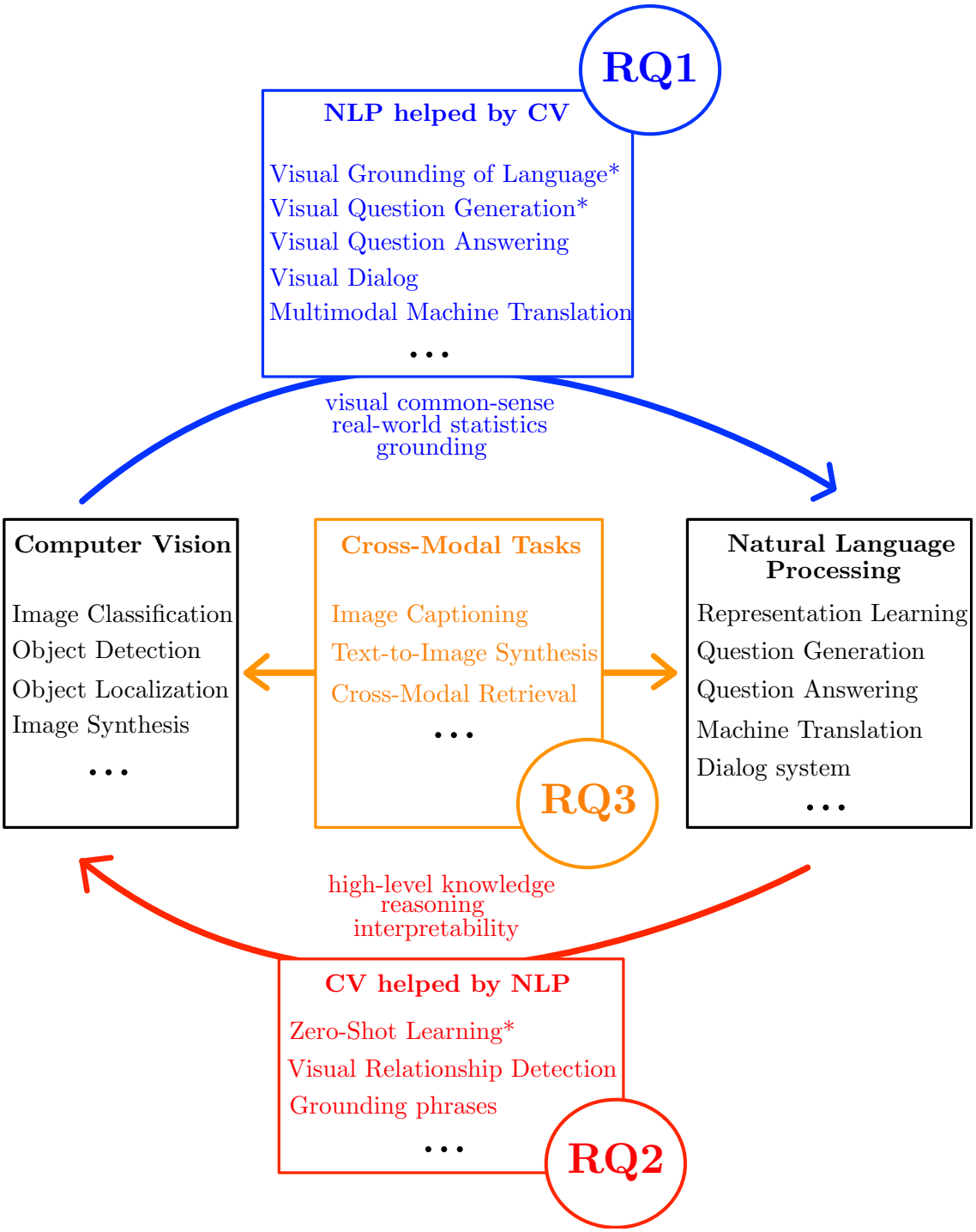

Figure 1.1 – Overview of the Research Questions and of the different multimodal tasks. * indicate tasks that are tackled as contributions in the present thesis.

in natural language are significantly different from their real-world frequency. The consequence is that real-world statistics derived from textual are significantly different to those derived from images. This can be explained intuitively: what is obvious (e.g, *bananas are yellow*, *humans have two legs*) is rarely stated in text, since it is supposed to be known by the human reader; the more an event is expected, the less it is likely to be conveyed.

From the aforementioned observations, we can infer that the textual and visual modalities present complementary characteristics, that may be efficiently leveraged in Machine Learning, for the mutual benefits of CV and NLP. This leads us to three Research Questions that will guide the present thesis:

- **RQ1: Can vision help to refine language understanding ?** Indeed, we can hint that vision can bring useful common-sense, grounding, and real-world statistics to NLP models.

  Since language is a purely human construction, often disconnected from a concrete reality, textual models that are trained only on textual data can lack crucial real-world information and make unrealistic predictions, such as "*the sky is green*" (Baroni 2016). Enhancing textual representations using complementary visual data is the purpose of Visual Grounding of Language. On the other hand, vision can be a useful complementary source of information for tasks that are commonly studied on text only, such as Question Answering (Weston et al. 2015) or Sentiment Analysis (Pang et al. 2007). For example, when asked to translate a sentence s from a language to another (Machine Translation), an image illustrating s may refine the translation prediction (Multimodal Machine Translation), by proposing a perceptual reference helping the model to desambiguate words.

  The capacity of vision to enhance language comprehension is a central topic of Chapter 4, Chapter 5 and Chapter 6. In Chapter 4, we investigate whether textual representations can benefit from visual information when text/image supervision is weak. In Chapter 5, we are interested in understanding to which extend vision can help Question Generation. In Chapter 6, we explore whether *sentence* representations can benefit from visual grounding, while most Visual Grounding of Language works consider *words*.

- **RQ2: Can language help to refine visual understanding ?** Language can serve to bring reasoning capabilities and high-level understanding of real-world scenes to CV models.

  Indeed, language can enhance the visual understanding capabilities of a model, especially when visual supervision is scarce, as in ZSL (Frome et al. 2013), which extends the traditional Image Classification tasks to classes that are unknown to the model. Indeed, traditional CV models tend to rely on a substantial amount of supervised data e.g., ImageNet (Deng et al. 2009) and MS COCO (T. Lin et al. 2014a), while it is possible to learn

high-quality textual representations in an unsupervised fashion. Thus, the natural supervision present in language may help CV systems to refine their recognition performances.

The capacity of language to enhance visual comprehension is an important issue in Chapter 3, Chapter 4 and Chapter 5. In Chapter 3 (resp. Chapter 5), we are interested in determining the presence (and, if possible, the nature) of visual knowledge present in word representations (resp. Neural Language Models). In Chapter 4, we are interested in tackling ZSL when the number of unknown classes is high.

- **RQ3: Can modalities be translated into one another ?** Navigating between modalities may shed some light on what makes them different, and bridge the gap between them.

  Historically, one of the first multimodal tasks was Cross-Modal Retrieval: from a given input (whether text or image), the goal is to find the element from the other modality that is semantically closest to the input. To do so, a shared multimodal space has to be built: in this vector space, visual and textual elements co-exist, and semantic distances can be computed between them. Methods to produce such spaces is an important problem in Multimodal Machine Learning: while first methods used statistical tools, such as Canonical Correlation Analysis (CCA) (Silberer et al. 2012), current approaches focus on learning local metrics, using, for example, triplet losses (Socher et al. 2014a).

  The more natural cross-modal direction is from images to text: indeed, it is intuitively easier to translate low-level pixels semantics into high-level abstraction, rather than the other way around. For example, Image Captioning (Karpathy et al. 2017) consists in generating a descriptive caption from an image. This task has been widely studied, in particular using the encoder-decoder framework (Sutskever et al. 2014): traditionally, the image is encoded in a vector using a CNN, and this vector is decoded sequentially into a sentence using a Recurrent Neural Network (RNN).

  The other direction, called Text-to-Image Synthesis (Gorti et al. 2018), is much more challenging. It builds upon the latest developments of Generative Adversarial Networks (Goodfellow et al. 2014), conditioned by a textual input, to generate images; due to the difficulty of the task, it has mostly been tackled within restricted visual domains, such as images of birds or flowers.

  The capacity of models to navigate betwen modalities is also a problem that we tackle in the present thesis, in Chapter 3, Chapter 4 and Chapter 5. In Chapter 3, we investigate how to make the predictions of cross-modal model more interpretable, by using specialized sub-models in a Bayesian framework. In Chapter 4, we tackle a traditional cross-modal task, Cross-Modal Retrieval, in a fully unsupervised setting, to know whether a latent

text/image supervision can be used to learn an alignment between modalities. Finally, in Chapter 5, we are interested in studying, when generating a question, the impact of using images, text or both simultaneously as input.

The inherent differences between language and vision is the challenge at the core of Multimodal Machine Learning, which is the subject of the present thesis. An overview is presented in Figure 1.1. Multimodal Machine Learning covers all tasks that require a joint understanding of language and vision. **RQ1** is at the origin of the first class of multimodal tasks: *NLP tasks helped by vision*. In these tasks, vision can either (i) refine language understanding, as in the Visual Grounding of Language (A. Lazaridou et al. 2015a) task, or (ii) extend standard NLP tasks to multimodal settings, as in Visual Question Answering (VQA) (Antol et al. 2015b), Visual Question Generation (VQG) (Y. Li et al. 2018) or Visual Dialog (Das et al. 2019). **RQ2** is at the origin of the second class of multimodal tasks: *CV tasks helped by language*. In these tasks, language brings reasoning capabilities and semantics to standard CV tasks, as in ZSL (Norouzi et al. 2014), which extends the Image Classification task, or Phrase Grounding (Karpathy et al. 2017), which extends the Object Detection task to phrases in natural language. **RQ3** is at the origin of the third class of multimodal tasks: *Cross-Modal Tasks*. In these tasks, modalities are translated into one another, such as in Image Captioning, Text-to-Image Synthesis or Cross-Modal Retrieval. These three classes are not mutually exclusive: for example, VQG extends Question Generation, a standard NLP task (*NLP helped by CV*) but it also corresponds to a *cross-modal* setting since an image is translated into a sentence.

1.3 Contributions and outline of the thesis

The contributions of the present thesis are outlined as follow.

In Chapter 2, we present existing works in Multimodal Machine Learning. We begin by describing standard uni-modal approaches to encode textual and visual information, as well as methods to fuse/merge uni-modal data to build multimodal representations. Then, we review three types of multimodal tasks: (1) *CV aided by NLP*: visual understanding tasks that benefit from textual knowledge, (2) *NLP aided by CV*: textual understanding tasks either refined using vision, or extended to a multimodal setting, and (3) *Cross-Modal tasks*: tasks where a modality is translated into another. In the following Chapters, we present novel contributions on various multi-modal tasks: ZSL, Transductive Zero-Shot Learning (T-ZSL), VQG, and Visual Grounding of Language, which cover the three groups of aforementioned multimodal tasks. The multimodal tasks of the present thesis are listed in Table 1.1.

In Chapter 3, we present a contribution in the field of ZSL. ZSL extends the image classification task to classes that are unknown to the model (unseen classes),

Chapter	Task	NLP aided by CV	CV aided by NLP	Cross-Modal
3	ZSL		✓	✓
4	T-ZSL		✓	✓
	CMR			✓
	VGL	✓		
5	VQG	✓		✓
6	VGL	✓		

Table 1.1 – Overview of the multimodal tasks covered in the present thesis. VGL hols for Visual Grounding of Language. CMR holds for Cross-Modal Retrieval.

thus relying on textual semantics brought by NLP models (*CV aided by NLP*); it also is a *Cross-Modal Task* as it consists in translating a visual input into a textual output (a class label). Our goal is to identify objects, delimited by bounding boxes, in images. As standard ZSL methods are solely based on the visual appearance of objects, we propose to use the visual context around objects to refine the predictions. To do so, we exploit semantic representations of class labels, and assume that they both contain information on the appearance of objects, and on the co-occurrence statistics with other objects in images. This work has been published: Eloi Zablocki*, Patrick Bordes*, Laure Soulier, Benjamin Piwowarski, and Patrick Gallinari (2019). "Context-Aware Zero-Shot Learning for Object Recognition". In: ICML 2019.

In Chapter 4, we present another contribution in ZSL. More precisely, we tackle Transductive Zero-Shot Learning (T-ZSL), which is also both a *CV aided by NLP* task and a *Cross-Modal Task*: in this setting, images corresponding to unseen classes are known to the model, but not their correspondence. We address a current limitation of T-ZSL models, that only consider datasets with a low number of unseen classes as they rely on the assumption that the visual space is well-clustered. As this hypothesis does not stand when the number of unseen classes is high (e.g, in the ImageNet dataset), we propose to align the textual and visual distributions with adversarial learning. Our model consists of a CycleGAN objective trained on unseen classes and a supervised objective trained on seen classes. We also evaluate our model on a Cross-Modal Retrieval task, and on Visual Grounding of Language. This work is currently under review at the Pattern Recognition journal: Patrick Bordes, Eloi Zablocki, Benjamin Piwowarski, and Patrick Gallinari "Transductive Zero-Shot Learning using Cross-Modal CycleGAN".

In Chapter 5, we present a contribution on VQG. VQG is an adaptation of a NLP task (Textual Generation) extended to a multimodal setting (*NLP aided by CV*); moreover, it can be seen as a *Cross-modal task* as, in the configuration where only the image is an input, the goal is to generate a textual output — VQG differs from Image Captioning in the sense that the textual output is a question, not a

description. Following recent works that showed the cross-lingual transferability of large Language Models such as BERT, our objective is to assess the *cross-modal* transferability of BERT. To do so, we (i) integrate visual data within BERT similarly to textual data and (ii) we apply BERT to VQG, which is the most convenient multimodal task as various inputs can be considered: purely visual, purely textual and multimodal. To generate a question, we propose *BERT-gen*, an extension of BERT able to generate an output based on uni- and/or multi-modal data. This work is currently under review at the EMNLP 2020 conference: Thomas Scialom*, Patrick Bordes*, Paul-Alexis Dray, Jacopo Staiano, Patrick Gallinari "BERT Can See out of the Box: On the Cross-Modal Transferability of Text Representations".

In Chapter 6, we present a contribution on the Visual Grounding of Language task. VGL aims at enhancing textual representations using visual information: it is thus a *NLP task aided by CV*. A large body of work learns multimodal word representations; however, the visual grounding of sentences remain under-explored. We determine the differences between sentences and words: contrarily to words, sentences can be visually ambiguous, carry non-visual information, or have a wide variety of paraphrases and related sentences. We derive from these assumptions objective functions, aimed at transfering visual information to sentence embeddings within an intermediate space, to avoid an over-constrained semantic space. This work has been published: Patrick Bordes*, Éloi Zablocki*, Laure Soulier, Benjamin Piwowarski, and Patrick Gallinari (2019). “Incorporating Visual Semantics into Sentence Representations within a Grounded Space”. In: EMNLP 2019.

Finally, in Chapter 7 we summarize the contributions of this thesis, and propose research perspectives for Multimodal Machine Learning.

CHAPTER 2

MULTIMODAL MACHINE LEARNING: BACKGROUND

Contents

Chapter abstract

In this Chapter, we give an overview of Multimodal Machine Learning, by presenting the main issues and by covering a variety of tasks that jointly leverage language and vision. We first describe techniques to represent textual data in Neural Language Processing (Section 2.1.1), and visual data in Computer Vision (Section 2.1.2). Then, we present various methods to integrate jointly textual and visual data within a multimodal framework (Section 2.1.3). We finally review previous works that tackle the three Research Questions discussed in the Introduction:

- ***RQ1: Can vision help to refine language understanding ?*** *In these works, the visual modality is used to help the interpretation of language in another modality (NLP aided by CV); for example, extensions of NLP tasks such as Visual Question Generation (VQG) (Y. Li et al. 2018), Visual Grounding of Language (Angeliki Lazaridou et al. 2015), Visual Question Answering (VQA) (Antol et al. 2015b), Visual Dialog (Das et al. 2019) or Multimodal Sentiment Analysis (Zadeh et al. 2016) (Section 2.2).*

- ***RQ2: Can language help to refine visual understanding ?*** *The textual modality is used to enhance visual reasoning and understanding (CV aided by NLP); for example, Zero-Shot Learning (Frome et al. 2013), Visual Relationship Detection (C. Lu et al. 2016b) (Section 2.3).*

- ***RQ3: Can modalities be translated into one another ?****: one modality is translated into another (Cross-Modal tasks), such as Image Captioning (Bernardi et al. 2016) or Text-to-Image generation (Gorti et al. 2018) (Section 2.4).*

2.1 From Mono- to Multi-modal Machine Learning

Traditional machine learning approaches rely on manually-designed data features. This *feature engineering* requires experts, is costly and generalizes poorly to new tasks. The paradigm of deep learning (LeCun et al. 2015) is different, and it is at the core of *representation learning*: the idea is to learn representations by leveraging large amounts of data, so that these representations encode meaningful semantic information and can be applied efficiently to downstream tasks.

In this section, we present representation learning techniques for uni-modal data. First, in the field of Natural Language Processing (Section 2.1.1), which aims at representing textual data. Then, in the field of Computer Vision (Section 2.1.2), for visual data. Finally, we show how to build multimodal representations from uni-modal data (Section 2.1.3).

2.1.1 Natural Language Processing

Natural Language Processing covers all techniques and models that process textual data — we present traditional NLP tasks in Table 2.1. Thus, Natural Language Processing (NLP) intends to capture the semantics, grammar and syntax of language into meaningful *representations*, and build models that learn such representations. In this section, we present techniques to learn textual representations from various granularities: words (Section 2.1.1.1), sentence (Section 2.1.1.2)

and documents (Section 2.1.1.3). In Section 2.1.1.4, we present language models, which are tools to estimate probability distributions over sequence of words.

Task	Description	References
Named Entity Recognition (NER)	Find and classify named entities in text into pre-defined categories, such as persons, organizations, locations, ...	(Guillaume Lample et al. 2016; Moreno et al. 2017)
Text summarization	Shorten a text into a summary representing the most imporant or relevant information of the original content	(Rush et al. 2015; Aries et al. 2019)
Question Answering	Answer questions posed by humans in a natural language	(Weston et al. 2015; D. Chen et al. 2017)
Machine translation	Translate a text from one language to another	(Bahdanau et al. 2014; Artetxe et al. 2018; Guillaume Lample et al. 2018a)
Dialog systems	Dialog with a human	(Vinyals et al. 2015a; Sordoni et al. 2015)
Sentiment analysis	Identify, extract, quantify affective states and subjective information	(Pang et al. 2007; Maas et al. 2011; Pontiki et al. 2016)
Parts-of-Speech (POS) tagging	Label a word in a text as a particular part of speech (*e.g.* noun, verb, adjective ...)	(Petrov et al. 2012; Nguyen et al. 2016)

Table 2.1 – **Examples of NLP tasks**. These tasks are conditioned by the quality of textual representations, that capture the meaning and semantics of textual data.

2.1.1.1 Word Representations

One-hot embeddings One-hot embedding is the simplest method to encode a word w. With a pre-defined dictionary of words $\mathcal{D} = \{w_1 ... w_N\}$, the one-hot embedding of w is defined as the vector $t_w = (\mathbb{1}_{w_i=w})_{i=1}^{N}$. Thus, this vector is sparse (all zeros except for a 1). However, semantics are not present in t_w. The geometrical cosine/inner-product between words that are semantically close, and the distance between words that are semantically different, is always the same: 0. These limitations led to the development of distributed word vectors, aimed at encoding the semantics of a word within reasonable dimensions.

Distributional Semantic Model (DSM) Distributional Semantic Models are based on the Distributional Hypothesis (Harris 1954), which states that *linguistic*

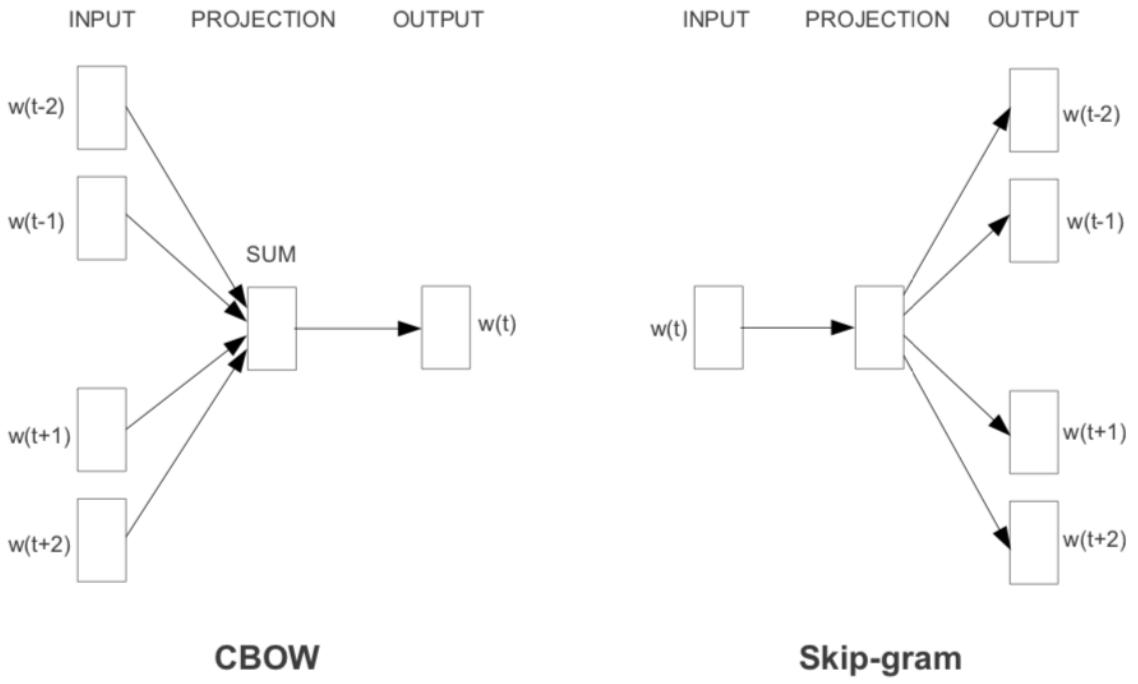

Figure 2.1 – Continuous Bag-of-Words (CBOW) and Skip-Gram Models. Illustration taken from Mikolov et al. 2013a.

items with similar distributions in textual corpora should have similar meanings — in other terms, *a word is characterized by the company it keeps* (Firth 1957). In practice, co-occurrence patterns of words in text are used to learn representations of word meaning, typically vectors t_w in a vector space of fixed dimension d, with d generally between 100 and 1000. The most used DSMs are Glove (Pennington et al. 2014), which relies on aggregated global word-word co-occurrence statistics from a textual corpus, and Word2vec (Mikolov et al. 2013b). Two objectives, illustrated in Figure 2.1, can be used in Word2vec:

- Skip-Gram (Mikolov et al. 2013b): For each word w_t occurring at position t, embedded by a vector (its Word2vec representation), the training objective is to predict the neighboring words in a text corpus, around a window of size $2c$:

$$\mathcal{L}_{skip-gram} = \frac{1}{T}\sum_{t=1}^{T}\sum_{-c\leq j\leq c, j\neq 0} \log p(w_{t+j}|w_t) \tag{2.1}$$

  The probability p is defined as: $p(e|c) \propto \exp(u_c^T.t_e)$, where u_c and t_e are two representations learned for each word depending on their role: for each word w, t_w is the Word2vec vector and u_w is the context vector.

- CBOW (Mikolov et al. 2013a): the training objective is to predict a word w_t given its context:

$$\mathcal{L}_{CBOW} = \frac{1}{T}\sum_{t=1}^{T} \log p(w_t | \sum_{-c \leq j \leq c, j \neq 0} w_{t+j}) \tag{2.2}$$

Several empirical observations can be made on DSMs:

- Levy et al. 2014 show that DSMs, like Glove and Word2vec, are equivalent to an implicit factorization of a word-word co-occurrence matrix; thus, they are variants of Latent Semantic Analysis (LSA), Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) algorithms.
- Words that share similar semantics are located in the same regions of the embedding space
- Some relations between words are linear. For example, Pennington et al. 2014 observe that "king queen = man woman", in the sense that $t_{king} - t_{queen} \approx t_{man} - t_{woman}$. Similarly, as illustrated in Figure 2.2, Mikolov et al. 2013b show that $t_{Madrid} - t_{Spain} \approx t_{Paris} - t_{France}$. More precisely, $t_{Madrid} - t_{Spain} - t_{Paris}$ is closer to t_{France} than to any other word.
- If two word spaces are built on two separate languages, using the same word representation model, then both spaces share similar structures [1]. This allows for efficient semi- or un-supervised alignment of word vectors between distinct languages, as done in bilingual lexicon induction such as Smith et al. 2017 and G. Lample et al. 2018.

Standard models such as Word2Vec and Glove map each word to a fixed point in a vector space, and the relationship between words is generally computed using cosine similarity. This does not account for the uncertainty of words (e.g., *food* is a broad concept that cover many aspects, whereas *rice* is more specific) and asymetrical relations between words, like inclusion (e.g., *Bach* is part of *composer*). To address this issue, Gaussian word embeddings (Vilnis et al. 2014) propose to represent words by *densities* learned in the space of Gaussian distributions.

ELMo (Peters et al. 2018) presents a new paradigm: *contextualized word representations*. Here, the embedding of a word in a sentence is a function of the entire sentence, not just the word itself. To do so, a Bidirectional Long-Short Term Memory (LSTM) (see Section 2.1.1.2) is trained on a large text corpus on a Language Model task. ELMo representations of each word are a linear combination of all internal layers of the LSTM for the corresponding token.

1. G. Lample et al. 2018 shows that it depends on the closeness between languages, e.g, French is closer to English than Czech

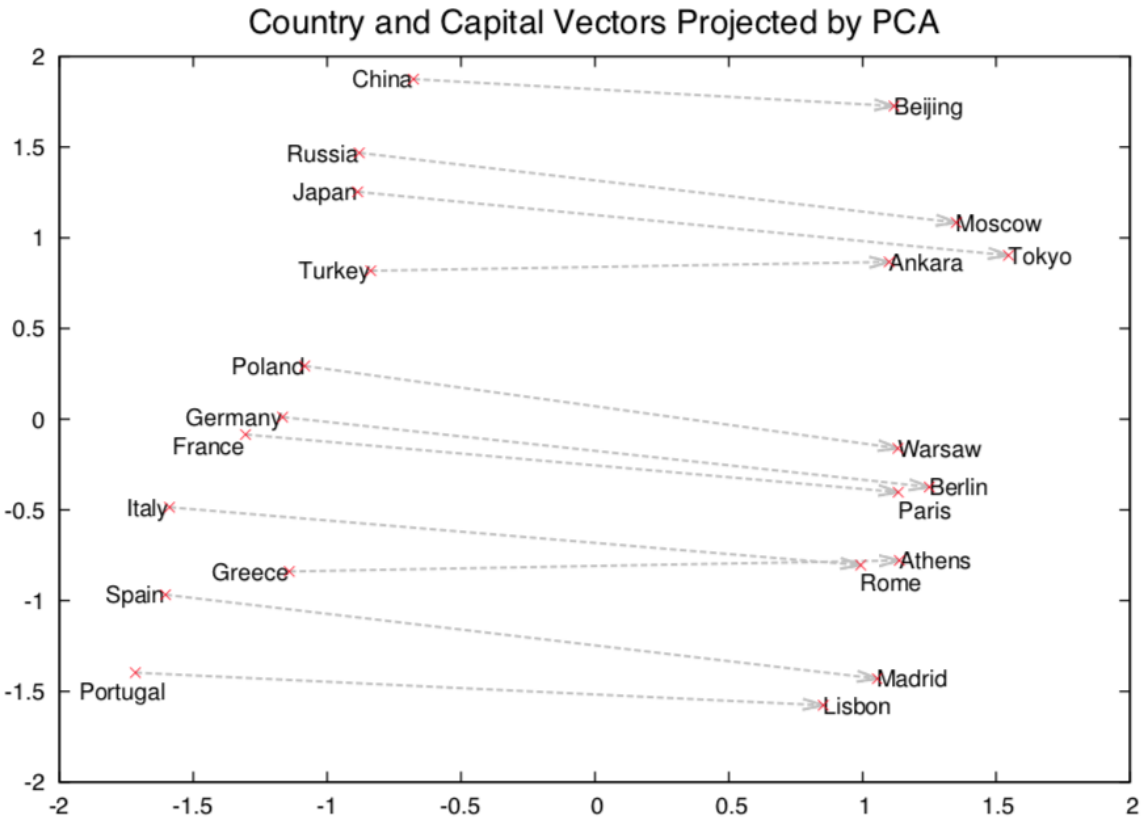

Figure 2.2 – Two-dimensional PCA projection of the 1000-dimensional Skip-gram vectors of countries and their capital cities. Illustration taken from Mikolov et al. 2013b.

Evaluation of word representations There exist various methods to evaluate word representations. On the one hand, unsupervised methods directly use word vectors to compute a score. On the other hand, supervised methods learn a classifier, generally evaluated by cross-validation. Examples are given in Table 2.2.

- **Semantic relatedness** (unsupervised method): determines whether the geometric distance between words given by the word model correlates with human judgements. In these tasks, we start from benchmarks annotated by humans — e.g, WordSim353 (Finkelstein et al. 2002), MEN (Bruni et al. 2014), SimLex-999 (F. Hill et al. 2015), SemSim and VisSim (Silberer et al. 2014) — where couple of words are given a similarity score between 0 and 10. For each couple of word, the predicted similarity score given by the word model is defined by the cosine similarity between both words. The semantic relatedness score is defined as the Spearman correlation between human ground-truth judgments and predicted similarities.

word 1	word 2	Similarity
football	tennis	6.63
stock	phone	1.62
coast	forest	3.15
jaguar	cat	7.42
journey	voyage	9.29
...	...	...

(a)

word	*animal*	*of metal*	*transportation*	...
bear	1	0	0	...
butterfly	1	0	0	...
scooter	0	1	1	...
shawl	0	0	0	...
wrench	0	1	0	...
...	...	...	...	...

(b)

word	Conc.
leopard	7.00
quiet	3.70
awareness	2.61
envelope	5.75
hound	6.83
....	...

(c)

word 1	word 2	word a	word b
Madrid	Spain	Paris	France
king	queen	man	woman
make	making	run	running
cat	kitten	dog	puppy
loose	lost	speak	spoken
...	...	...	...

(d)

Table 2.2 – Word evaluation benchmarks: (a) Semantic relatedness; examples from WordSim353 (Finkelstein et al. 2002). (b) Feature-norm prediction; examples from McRae et al. 2005. (c) Concreteness prediction; examples from USF (Nelson et al. 2004). (d) Analogy prediction.

- **Analogy prediction**: It determines whether relations between words in language can be found in the vector space, e.g, if $t_{king} - t_{queen} \approx t_{man} - t_{woman}$.

- **Concreteness prediction** (supervised method): The goal is to predict to which extent a word is abstract or concrete (ground-truth scores have been made by human annotators) given the word representations. This is done on the USF dataset (Nelson et al. 2004) (3260 English words).

- **Feature-norm prediction** (supervised): The goal is to predict characteristics of objects (e.g, *has legs*, *is green*) given their word representation. There are 417 entities, with a total of 43 characteristics divided into 9 categories (*taste*, *sound*, *tactile*, *color*, etc.).

For contextualized word embedding models like ELMo, evaluations are made by integrating ELMo embeddings to pre-existing NLP models, on tasks such as Question Answering on the Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al. 2016), or Named Entity Recognition (NER) on the CoNLL 2003 NER task (Sang et al. 2003).

2.1.1.2 Sentence representations

Architectures Over the last years, several approaches have been proposed to learn semantic representations for sentences. The most naive approach is the Bag-of-Words (BoW) model: a sentence $s = w_1...w_n$ is represented as a sum (or weighted average) of the one-hot embeddings if its words. This approach has the same limitations of one-hot encodings (no encoded semantics); in addition, the information of word order is not encoded. This is a problem for sentences, as word order can be determinant, e.g, *Lee Harvey Oswald assassinated JFK* and *JFK assassinated Lee Harvey Oswald* have entirely different meanings. Neural approaches have been proposed to tackle these issues.

Since sentences are intrinsically sequential, a tool to leverage sentences are Recurrent Neural Network (RNN). Unlike classical neural networks, which assume the inputs to be independent, RNNs make use of sequential data and perform the same task on every word of a sentence $x_1...x_T$. The key element is the hidden state h_t, which acts as a memory and is computed based on the input at the current state x_t and the previous state h_{t-1}:

$$h_t = f(Ux_t + Wh_t) \tag{2.3}$$

The output o_t is calculated with the hidden state: $o_t = \sigma(V_o.h_t + b_o)$. The sentence embedding correponds to the last hidden state h_T.

Standard RNNs suffer from limitations: as the gap between the relevant information and the point where it is needed grows, RNNs become unable to connect the information due to the vanishing gradient problem (Y. Bengio et al. 1994). To tackle this issue, LSTM networks (Hochreiter et al. 1997) have been designed. The LSTM is a RNN aimed at handling long-term dependencies. The key element of the LSTM is the cell state c_t, which acts like a conveyor belt flowing long-term information. The network can remove or add information to the cell state using:

- a forget gate, deciding which information is to be thrown away from the cell state: $f_t = \sigma(W_f.[h_{t-1}, x_t] + b_f)$
- an input gate, deciding which information is to be stored in the cell state: $i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i)$
  This gate also generates candidate values, that could be added to the state: $C_t^{'} = \sigma(W_C.[h_{t-1}, x_t] + b_C)$
- an output gate, deciding which part of the cell state to output: $o_t = \sigma(W_o.[h_{t-1}, x_t] + b_o)$

The cell state c_t is updated using a combination of its past state c_{t-1} (moderated by the forget gate f_t) and the candidate state $c_t^{'}$ (moderated by the forget gate i_t): $c_t = f_t.c_{t-1} + i_t.c_t^{'}$

To simplify the complex structure of the LSTM, a more efficient architecture has been proposed, that does not rely on a memory unit: the Gated Recurrent Unit (GRU) (Cho et al. 2014). This network combines the forget and input gates into an *update gate* and merges the hidden state and the cell state. The next hidden state is a linear combination of the previous hidden state and the state update:

$$h^t = (1 - z^t) \bigodot h^{t-1} + z^t \bigodot \overline{h}^t \tag{2.4}$$

where the update gate is $z^t = \sigma(W_z[h^{t-1}, x^t])$. The state update is defined by: $\overline{h}^t = \tanh(W[r^t \odot h^{t-1}, x^t]))$, where r^t is a *reset* gate: $r^t = \sigma(W_r[h^{t-1}, x^t])$.

Since traditional RNNs read words from right to left, the representation of the sentence tends to forget the beginning of the sentence. Following this consideration, bidirectional RNNs (Schuster et al. 1997) aim at learning a symmetric sentence representation, able to incorporate information from both the beginning and the end. For a sequence of T words, a bidirectional RNN computes a set of T vectors h_t, which is the concatenation of a forward RNN and a backward RNN that read the sentences in two opposite directions.

The Transformer (Vaswani et al. 2017) proposes a different paradigm than the RNN: using only attention mechanisms, and discarding recurrent and convolutional techniques entirely. The motivation is twofold: first, attention has proven useful to model long-term dependencies (Bahdanau et al. 2014; Y. Kim et al. 2017) (a long-standing issue for RNNs); second, the Transformer enables efficient parallelization possibilities for training, unlike RNNs that are inherently sequential.

The Transformer follows the encoder-decoder framework (Sutskever et al. 2014): it is composed of a series of encoding layers and a series of decoding layers, with each layer having independent weights. The novelty and the fundamental building block of the Transformer is the *self-attention layer*: its goal is to encode each word with an attention over all tokens of the previous layer. The self-attention layer takes n word vectors as input, and outputs n word vectors, that are written as linear combinations of a transformation of the input vectors.

In the Transformer, each encoding layer is composed of: (i) a self-attention layer and (ii) a feed-forward neural network applied to each token vector of the sequence. The encoder output is used at each decoding time step in the *encoder-decoder attention* blocks. Each decoder layer is composed of (i) a self-attention layer, (ii) an encoder-decoder attention — helping the decoder to focus on relevant parts of the input sentence — and (iii) a feed-forward network. At each decoding step, the input of the decoder are the target tokens decoded up to the current step. The first decoder input is a special token "beginning of sentence"; at step k, there are k tokens $w_1, ...w_k$ as input: the output of the decoder at step k is taken as the next token w_{k+1}. The process stops when a special token "end of sentence" is encountered.

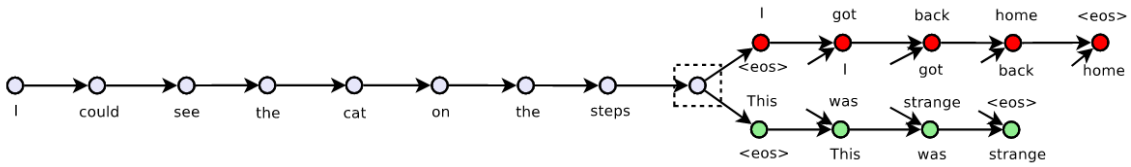

Figure 2.3 – Skip-Thought objective. Illustration taken from Kiros et al. 2015.

Learning procedures Historically, first methods relied on *supervised* approaches to learn sentence representations. In supervised techniques, the principle is to apply sentence representations to a supervised task, using varying architectures such as recursive networks (Socher et al. 2013), convolutional networks (Kalchbrenner et al. 2014) or self-attentive networks (Zhouhan Lin et al. 2017). The common limitation of such methods, as pointed out in (Kiros et al. 2015), is that it produces highly task-dependent sentence vectors; for this reason, we will focus on *unsupervised* techniques in the following.

Unsupervised approaches, by benefiting from large amount of training data, tend to produce universal and task-independent representations: FastSent (Felix Hill et al. 2016), Word Information Series (Arroyo-Fernández et al. 2019), Universal Sentence Encoder (D. Cer et al. 2018) are notable examples.

SkipThought (Kiros et al. 2015) — illustrated in Figure 2.3 — proposes an extension of the distributional hypothesis at the sentence level: *sentences that appear in similar places in text tend to have similar meanings*. Thus, the training objective aims, for each sentence, at predicting the neighboring sentences in a vast text corpus: the Toronto BookCorpus (Yukun Zhu et al. 2015), composed of 11K unpublished books and about 71M sentences. It extends the Word2Vec principle to the sentence level: the encoder-decoder tries to reconstruct the surrounding sentences of an encoded passage. The authors use a RNN encoder with GRU activations and an RNN decoder with a GRU.

More recently, the QuickThought (Logeswaran et al. 2018) model, instead of training an encoder-decoder to predict neighboring sentences, adopts a different approach: a sentence representation is trained to select the adjacent sentence among three candidate sentences, among which two are negative samples.

Evaluation of sentence representations To evaluate sentence representations, two types of evaluations exist:

- **Semantic relatedness**: two benchmarks are used: Semantic Textual Similarity (STS) (D. M. Cer et al. 2017) and Sentences Involving Compositional Knowledge (SICK) (Marelli et al. 2014a) — examples given in Table 2.5. They both consist of pairs of sentences that are associated with human-labeled similarity scores. STS is subdivided into three textual sources: *Captions* contain concrete sentences describing daily-life actions, whereas the others contain

Positive	Negative
a taut , intelligent psychological drama	might best be enjoyed as a daytime soaper
an imaginative comedy/thriller	plodding , peevish and gimmicky
a fascinating and fun film	a soulless , stupid sequel . . .

Table 2.3 – MR examples

Positive	Negative
"I still don't think it is a trade secret," he said yesterday.	Dusty had battled kidney cancer for more than a year.
"We don't think that there is a trade secret here."	Dusty had surgery for cancer and had a kidney removed.

Table 2.4 – MSRP examples

more abstract sentences: news headlines in *News* and posts from user forums in *Forum*.The Spearman correlations are measured between the cosine similarity of our learned sentence embeddings and human-labeled scores.

- **Classification**: a logistic regression classifier is learned from the extracted sentence embeddings, and the classification accuracy is reported. The tasks are the following: Multi-Perspective Question Answering (MPQA) (Wiebe et al. 2005), Movie Review (MR) (Pang et al. 2005) — examples given in Table 2.3 — , Subjectivity/Objectivity (SUBJ) (Pang et al. 2004), Customer Reviews (CR) (M. Hu et al. 2004), binary sentiment analysis on Stanford Sentiment Treebank (SST) (Socher et al. 2013), paraphrase identification on Microsoft Research Paraphrase (MSRP) (Dolan et al. 2004) — examples given in Table 2.4 — as well as two entailment classification benchmarks: Stanford Natural Language Inference (SNLI) (Bowman et al. 2015) and SICK (Marelli et al. 2014b).

2.1.1.3 Document Representation

Going beyond the sentence level, *document* representations have been learned. This granularity level is the most challenging, as documents are aggregates of sentences, that can present different semantics and themes.

Sentence 1	Sentence 2	Similarity
A black and brown cat is eyeing a fly	The man is eating cereal	1
A dog is emerging from a lake	An animal is emerging from a lake	4.6
A dog is running through the snow	No dog is running through the snow	3.1
A kid is splashing in the pool	A kid is splashing in the ocean	4.1
Two children are playing	A kid is splashing in the ocean	2

Table 2.5 – SICK examples

Historically, Term Frequency-Inverse Document Frequency (TF-IDF) has been used to represent documents in the Information Retrieval field (Salton et al. 1984; Jones 2004; Vulic et al. 2015). TF-IDF aims at measuring how important is a word to a document in a collection of corpus. It is composed of (i) a term frequency (TF) — measuring the number of occurrences of a word in a document — multiplied by (ii) the logarithm of the inverse document frequency (IDF) — measuring how much information the word provides, IDF is the number of documents in the corpus containing that word. Intuitively, common words like *the* will have a high TF but a very low IDF since it is present in all documents. Thus, a high TF-IDF means that the word is representative of the given document. Once the TF-IDF of all words in all documents have been computed, a vocabulary is fixed and each document is represented by a vector containing their TF-IDF values for each word of the vocabulary. The limitation of this approach is that the order of words is not considered; it also does not encode semantics.

To solve this issue, the ParagraphVector (Le et al. 2014) model has been proposed: it is close to the Word2Vec model (Mikolov et al. 2013b) and is frequently used to encode the semantics of variable-length documents. The training objective is: (i) to predict a word given its context, (ii) to predict words that appear in a small window by leveraging a vector representation of the document. Unlike other works, here, sentences are seen as basic units of documents, whereas sentences are usually seen as a composition of words. (Le et al. 2014) extends the CBOW idea and learns fixed-length feature representations by introducing a distributed sentence indicator. Each sentence is represented by a dense vector which is trained to predict words in the document. The downside of this method is that the sentence indicator has to be estimated at test time.

2.1.1.4 Language Models

Language Models are models that assign probabilities to sequences of words $P(w_1...w_n)$. Their goal is not directly to learn representations, but since the 2000s, they often rely on textual representations (in general, word embeddings). The foundation of language models is the following equation:

$$P(w_1...w_n) = \prod_{i=1}^{n} P(w_i|w_1...w_{i-1}) \tag{2.5}$$

N-gram A *n-gram* is defined as a sequence of n consecutive words. N-gram models use a Markovian assumption, by supposing that the words necessary to predict w_i are the n last words (Y. Bengio et al. 2003). Given this assumption, Equation 2.5 can be re-written as:

$$P(w_1...w_n) = \prod_{i=1}^{n} P(w_i|w_{n-1-i}...w_{i-1}) \tag{2.6}$$

The value of $P(w_i|w_{n-1-i}...w_{i-1})$ can be estimated by the number of co-occurrences of the n-gram $w_{n-1-i}...w_{i-1}, w_i$ divided by the number of co-occurrences of $w_{n-1-i}...w_{i-1}$, where co-occurrences are computed on vast text corpus. The n-gram model considers word order, and is simple to implement since it only involves counting. However, it faces several limitations: (i) when n is high, the number of samples to gather for estimating reliably the distribution is high; (ii) some particular n-grams may not be present in text corpus, thus leading to a probability of 0 — this problem is alleviated with a smoothing techniques (Manning et al. 2008).

Neural Language Model To alleviate the issues of the n-gram model, the Neural Language Model (NLM) has been proposed.

Historically, RNNs (Y. Bengio et al. 2003) have been used to model word co-occurrence probabilities, by using continuous representations of words. The hidden state h_i at time i aims at synthesizing previous history of the sequence. Thus, NLMs rely on the following estimation:

$$P(w_{i+1}|w_1...w_i) \propto \exp(V_o.h_i + b_o) \tag{2.7}$$

More recently, following the success of Transformers, NLMs have relied on Transformer architectures, in particular with OpenAI GPT (Radford et al. 2018) and BERT (Devlin et al. 2019). Moreover, for these models, the paradigm is different: unlike *feature-based approaches* (e.g., SkipThought), which learn textual representations that are used directly (without fine-tuning) on task-specific architectures, these model use a *fine-tuning* approach: they introduces minimal task-specific parameters, and fine-tune all model parameters on downstream tasks.

In OpenAI GPT (Radford et al. 2018), a Transformer is trained to predict the next token; each token can only attend to previous tokens. However, this approach faces an important limitation, as its left-to-right architecture prevents tokens from attending to future tokens.

BERT (Devlin et al. 2019) — illustrated in Figure 2.4 — alleviates the problem of the uni-directionality of OpenAI GPT by proposing a new objective called Masked Language Model (MLM). Under MLM, some words, that are randomly selected, are masked; the training objective aims at predicting them. The authors also train their model on a next sentence prediction task. In BERT, words are embedded within an Embedding Layer, that converts them into vectors of a fixed dimension d (here, $d = 768$ for BERT-BASE and $d = 1024$ for BERT-LARGE). As illustrated in Figure 2.5, each word is embedded by the sum of (i) a token embedding, (ii) a segment embedding, indicating in which sentence the word is, (iii) a position embedding, encoding the position of the word in the input sequence — indeed, as BERT is a Transformer, the order information might be lost without this token. These vectors are fed as input of a 12-layer Transformer.

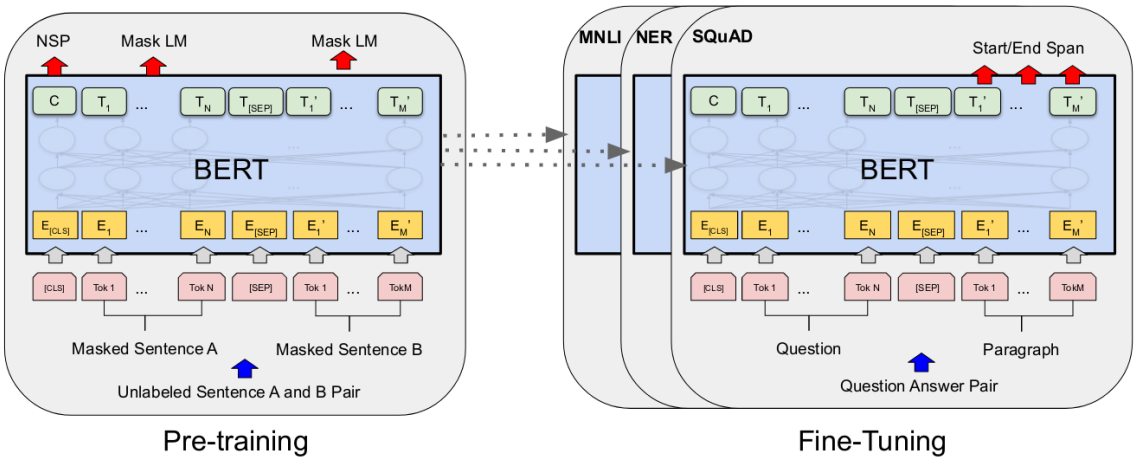

Figure 2.4 – BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. Illustration taken from Devlin et al. 2019.

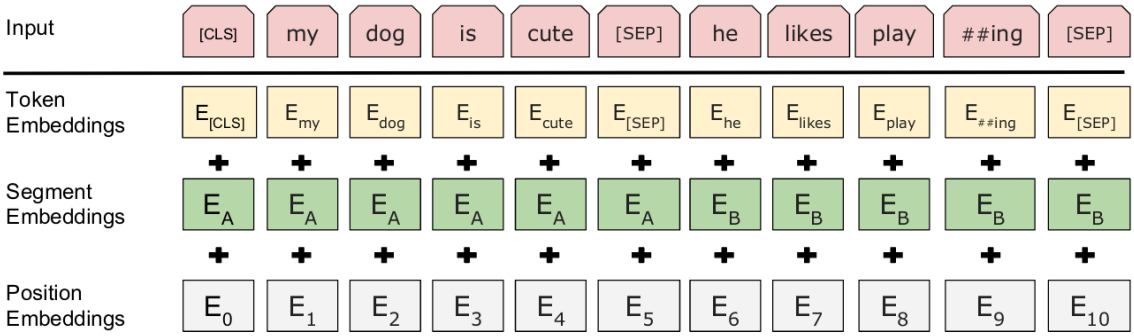

Figure 2.5 – BERT input representations. Illustration taken from Devlin et al. 2019.

Furthermore, the authors propose to use two special tokens: [SEP] separates sentences in the input sequence, and [CLS] is append at the beginning of the input sequence, and constitutes a global sequence embedding, that is often fed as input for discriminative downstream tasks.

2.1.2 Computer Vision

Computer Vision covers a variety of tasks that deal with the understanding of visual data — throughout this thesis, we only consider data from images. Computer Vision encompasses a variety of tasks, some of which are listed in Table 2.6.

To deal with these tasks, it is necessary to represent the images in meaningful semantic spaces: this is the focus of the present section. We present traditional hand-crafted features for images in Section 2.1.2.1, and deep representations with Convolutional Neural Networks in Section 2.1.2.2.

Task	Description	References
Image classification	Determine the right class of an image among a pre-defined set of labels	Deng et al. 2009; Krizhevsky et al. 2012; Szegedy et al. 2016a
Object Localization	Determine the position of a given object in an image	Uijlings et al. 2013; Girshick et al. 2014
Object Detection	Finding instances of objects in an image	Sermanet et al. 2014; S. Ren et al. 2017; Redmon et al. 2016
Semantic Segmentation	Label each pixel in the image by a category label	Hariharan et al. 2014; J. Long et al. 2015; Hariharan et al. 2015
Image Synthesis	Generating targeted modifications of existing images or entirely new images	Radford et al. 2016; Oord et al. 2016; J. Zhu et al. 2017
Image Colorization	Converting a grayscale image to a full color image	R. Zhang et al. 2016; Cheng et al. 2015

Table 2.6 – **Examples of Computer Vision tasks**. These tasks are conditioned by the quality of visual representations, that capture the meaning and semantics of visual data.

2.1.2.1 Hand-crafted Features

Before the rise of Convolutional Neural Networks, images were processed using hand-crafted features, i.e. features obtained without a learning phase, and that are used with standard Machine Learning (ML) models, e.g, naive Bayes or Support Vector Machine (SVM) (Tong et al. 2001). Since pixel information is rarely semantically relevant, and necessitates huge memory storage, hand-crafted approaches aim at finding local features that are later aggregated to build global image features.

The first step to represent an image is *feature detection*, i.e. detecting image keypoints, or salient regions of an image (often found at edges and changes of color intensity). The second step, *feature description*, aims at generating a vector able to represent these features. An important challenge is to produce features that are robust to the various appearances an object can have in images, due to the change of luminosity, rotation of the camera, distortion of the object, etc. A well-known effort toward that direction is the Scale-Invariant Feature Transform (SIFT) (Lowe 2004) model. The goal of SIFT is to produce visual features that are robust to object distortions, intensity changes, and point of view. To do so, images undergo convolutions of various Gaussian kernels at different scales: the produced SIFT

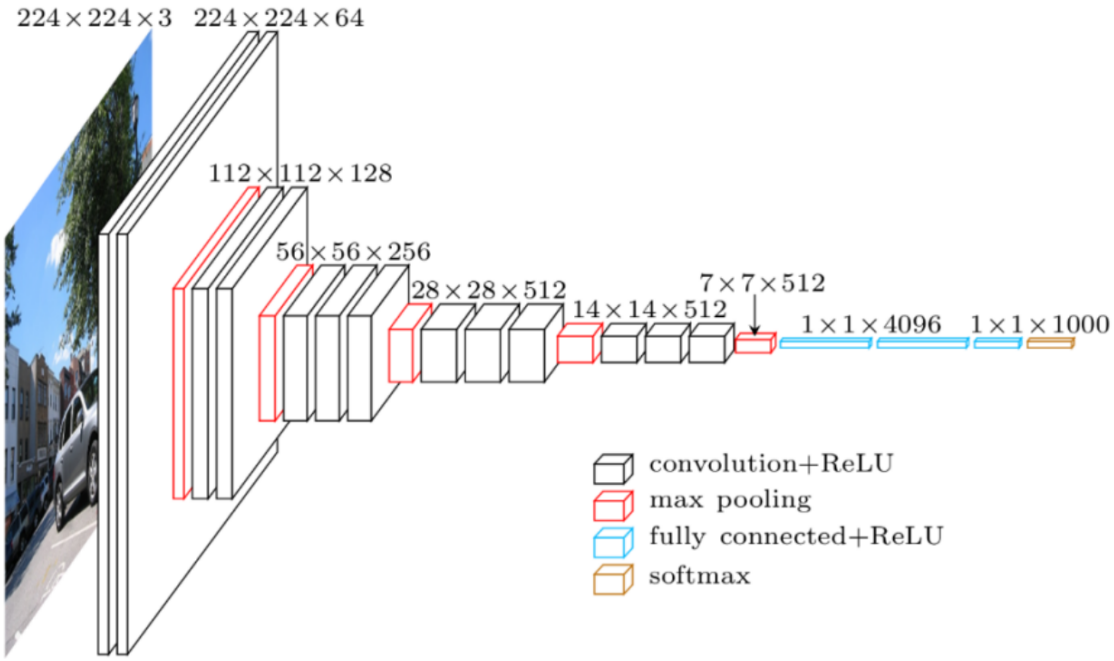

Figure 2.6 – Convolutional Neural Network. In that case, a VGG network (Simonyan et al. 2014). Illustration taken from Durand 2017.

features are invariant to scaling and rotation transformations. Histogram of Oriented Gradient (HOG) (Dalal et al. 2005) is another local feature descriptor. In this technique, occurrences of gradient orientations are counted in localized portions of an image. Unlike SIFT, the descriptors are computed on a dense grid of cells that are uniformly spaced, and the accuracy is further improved by using ovelapping local constrast normalization.

The third step is *codebook generation*. In this step, all feature descriptors are fed into an unsupervised clusterization algorithm, e.g, k-means clustering, to generate codewords (i.e. visual objects).

In the final step, images are represented using a Bag of Visual Words (BoVW) model (Qiu 2002). Thus, an image is embedded as a sparse vector with 1s at the positions of visual words. Hoewever, as in the BoW model for sentences, the structure of the image is ignored, whereas the position of objects in an image is paramount to understand visual semantics. Moreover, the design of feature descriptors is often task-dependent, i.e. filter banks include expert knowledge (Y. M. Lu et al. 2007). Hand-crafted methods thus require human intervention, and produce visual features that do not generalize to other tasks.

2.1.2.2 Convolutional Neural Networks

The Convolutional Neural Network (CNN) (Fukushima et al. 1982) is a famous example of the success of deep learning over hand-crafted features. The paradigm

is thus to *learn* high-quality image features that are not task-specific, and that can be applied to downstream tasks. A CNN is a hierarchical neural network, and aims at learning high-level abstractions by stacking a series of layers: the first layers detect corners and edges, whereas the final layers can recognize precise object (e.g, a face, a dog, a cat). With CNNs, the feature extraction phase is shunned: the input is the image and, when the CNN is trained on a classification task, the output is a distribution of probability over a set of visual objects fixed in advance.

The basic CNN architecture (illustrated in Figure 2.6) includes four basic components:

- *Convolution layer*: In this layer, various filters are used to convolve the whole image, producing a series of *feature maps*. The convolution operator has useful properties in the case of images: commutativity, associativity, distributivity and multiplicative identity. As noted in (Zeiler et al. 2014), the convolution operation (i) enables a reduced number of parameters due to weight sharing in a feature map, (ii) learns correlations among neighboring pixels, (iii) produces invariant features toward the location of the object

- *Non-linear activation layer*: following the discovery of the limitations of the *tanh* and sigmoid activation functions, that lead to a vanishing gradient problem, the Rectified Linear Unit $ReLU(x) = max(x, 0)$ is commonly used in current CNNs, as it solves the vanishing gradient problem and allows for a better and faster training

- *Pooling Layer*: this layer reduces the number of spatial dimensions of the feature maps

- *Fully-connected Layer*: it is the last layer of the CNN, that takes as input the output of the final pooling layer and outputs a probability distribution over the visual classes of the dataset.

The success of CNNs over the last decade can be attributed to several factors:

- the back-propagation algorithm, proposed in LeCun et al. 1989. This algorithm calculates the gradient of the loss function by applying the chain rule, thus computing the gradient one layer at a time, starting from the last layer, and avoiding redundant calculations

- efficient regularization techniques, such as drop-out (Nitish et al. 2014), in which some nodes are randomly ignored during training to avoid over-fitting

- the creation of large-scale image datasets such as ImageNet (Deng et al. 2009) (1.4M images corresponding to 1K classes), Visual Genome (Krishna et al. 2017) (100K images with fine-grained annotations) and MS COCO T. Lin et al. 2014a (180K training images with 5 corresponding captions each)

- the exponential increase of computing power with Graphics Computing Units (GPUs) allowing for fast convolution and matrix multiplication operations.

The architecture of CNNs has evolved into integrating more and more layers in the last decade. The first CNNS, such as LeNet (Lecun et al. 1998) and AlexNet (Krizhevsky et al. 2012) (first CNN architecture to even win the ImageNet challenge) had respectively 8 and 7 layers, whereas more recent CNNs such as VGG (Simonyan et al. 2014) and GoogleNet (Szegedy et al. 2015) have respectively 16 and 19 layers. The development of new architectures for visual tasks is still active. For example, using *residual layers*, CNNs like ResNet (K. He et al. 2016) can learn even deeper networks (152 layers). In the latest CNN developments, a Neural Architecture Search Network (NASNet) (Zoph et al. 2018) is used to determine the optimal architecture via reinforcement learning, leading to a smaller model size and lower complexity.

2.1.3 Building multimodal representations from mono-modal representations

In Section 2.1.1, we presented standard ways to produce textual representations, and in Section 2.1.2, standard ways to produce visual representations. In the multimodal tasks that we present in the rest of this Chapter, both textual *and* visual information have to be leveraged jointly. Thus, there is a need to build multimodal representations possibly from uni-modal representations: this is the purpose of the present section.

In Section 2.1.3.1, we present *fusion* techniques, where information from both modalities has to be aggregated into a unique vector, generally to be decoded into an output. This is useful in tasks like Visual Question Answering (VQA) (Section 2.2.2.2) or Multimodal Machine Translation (MMT) (Section 2.2.2.4), that have both a textual and a visual input.

In Section 2.1.3.2, we present techniques to build a *shared multimodal space*. This is often used in tasks such as Cross-Modal Retrieval (Section 2.4.4) or Zero-Shot Learning (ZSL) (Section 2.3.1), where images and texts are projected into a common representation space, in which similarities can be computed between elements from both modalities.

Finally, in Section 2.1.3.3, we present a recent approach to jointly integrate textual and visual information: Multimodal NLMs, that follow the success of BERT (Devlin et al. 2019). This approach is different from the *fusion* and *shared* as it is not oriented toward learning representations, but rather extends the NLM principle to multimodal data.

2.1.3.1 Fusion

Multimodal fusion aims at producing a multimodal representation when given a textual and a visual representation as input. A multimodal fusion function is thus a function f_θ, parameterized by θ, which takes as input a textual representation t (of dimension d_t) and a visual representation v (of dimension d_v) and outputs a vector $f_\theta(t, v)$. The problematic of multimodal fusion is at the core of the VQA task (see Section 2.2.2.2). Indeed, in VQA, the input is an image (visual) and a question (textual), and the model has to process these two elements to choose an answer.

Concatenation The simplest way to combine textual and visual information is to concatenate textual and visual representations:

$$f_\theta(t, v) = t \oplus v \tag{2.8}$$

where $\oplus$ designates the concatenation operator (in that case, $\theta = \emptyset$). This approach is used, for example, in the GroundSent model (Kiela et al. 2018) (see Section 2.2.1 for more detail)— a baseline model for one of our contributions (see Chapter 6) —, in which the final multimodal sentence representation is the concatenation of a purely-textual SkipThought vector (obtained from textual data) and a grounded one (obtained with an Image Captioning dataset).

To extract more meaningful representations, some methods learn a Multi-Layer Perceptron (MLP) — of parameters θ — on top of the concatenated vector:

$$f_\theta(t, v) = MLP_\theta(t \oplus v) \tag{2.9}$$

This method is used, for example, in the MDN-VQG model (Patro et al. 2018a) for Visual Question Generation — a baseline model for one of our contributions (see Chapter 5) — where a caption embedding vector and an image embedding vector are fusioned using Equation 2.9. The resulting multimodal vector is used to condition a decoder module that produces a question.

Element-wise product t and v can also be combined using a simple element-wise product, which suggests that $d_v = d_t$:

$$f_\theta(t, v) = t \odot v \tag{2.10}$$

where $\odot$ designates the element-wise product operation.

This technique is used in some VQA works (Antol et al. 2015a; J. Kim et al. 2016; R. Li et al. 2016), where v is the image vector and t the question vector; $f_\theta(t, v)$ is usually used as input of a classifier to determine an answer.

Bilinear models To allow for more complex interactions to occur between modalities, t and v can also be fusioned using a tensor $T_\theta \in \mathbb{R}^{d_t \times d_m \times d_v}$:

$$f_\theta(t, v) = t.T_\theta.v \tag{2.11}$$

This is often done in VQA (see Section 2.2.2.2) (Fukui et al. 2016; J. Kim et al. 2017). Due to the high dimension of the tensor T_θ, some approaches attempt to simplify the learning of T_θ by decreasing the number of parameters, for example with Tucker decomposition techniques (Ben-younes et al. 2017) or a stack of low-rank matrice (Z. Yu et al. 2017).

2.1.3.2 Shared spaces

In *shared spaces* approaches, images and texts are mapped to a common representation space. In this space, comparisons can be made between elements of distinct modalities (Weston et al. 2010; J. Wu et al. 2017): thus, this is a fundamental aspect of multimodal machine learning. Such a space can enable, given queries from one modality, to retrieve elements from another modality via a nearest neighbor search. This is commonly the case in Zero-Shot Learning (see Section 2.3.1), where a shared space is learned and, given an image, the class label which is closest to the projected image is retrieved. Learning meaningful shared spaces is at the core of the Cross-Modal Retrieval field (see Section 2.4.4).

Global alignment methods In global alignment methods, two mappings are learned to map the textual and visual space to a joint space, so that regions that are semantically similar across modalities are mapped closely in the common representation space.

Historically, Canonical Correlation Analysis (CCA) was one of the first used methods (Silberer et al. 2012; Gong et al. 2014). CCA aims at finding linear combinations between two sets of observations which have maximum correlation with each other. It also reduces the dimensionality of textual and visual data while preserving the most important interactions between them.

More formally, let $T \in \mathbb{R}^{N \times d_t}$ a matrix of N textual observations, and $V \in \mathbb{R}^{N \times d_v}$ the matrix of the corresponding visual observations. The goal of CCA is to find two projection matrices $P_t \in \mathbb{R}^{n \times d_v}$ and $P_v \in \mathbb{R}^{n \times d_v}$ (with $n \leq min(d_v, d_t)$), such that the original data T and V can be projected in a common space of dimension n with $T.P_t^T$ and $V.P_v^T$. This is done by first finding two vectors $t_1 \in \mathbb{R}^{d_t}$ and $v_1 \in \mathbb{R}^{d_v}$ that maximize the correlation $\rho(x, y) = corr(T.x^T, V.y^T)$. $T.t_1^T$ and $V.v_1^T$ are called the first pair of canonical variables. Then, two new vectors t_1 and v_1 that maximize ρ subject to the constraint that they are to be uncorrelated with the first pair of canonical variables; they are called second pair of canonical variables. The procedure is repeated n times and the $(t_i)_{i=1}^n$ (resp the $(v_i)_{i=1}^n$) are the rows of P_t (resp. P_v).

CCA has been used in many multimodal tasks; for example, in Visual Grounding of Language (see Section 2.2.1) (Silberer et al. 2012; Silberer et al. 2013; Felix Hill

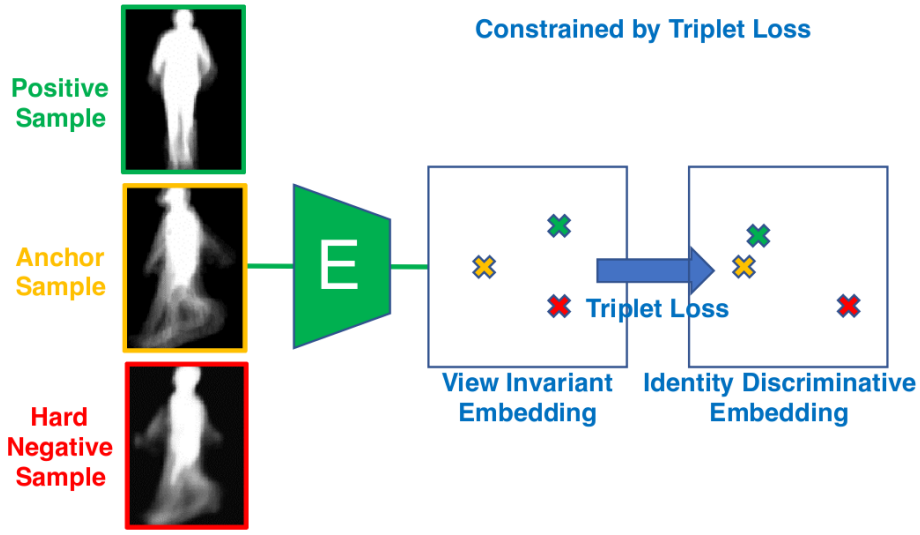

Figure 2.7 – Triplet Loss. Illustration taken from B. Hu et al. 2018.

et al. 2014b; Loeub et al. 2016) and Cross-Modal Retrieval (see Section 2.4.4) (Andrew et al. 2013; F. Feng et al. 2014).

Local metric learning methods Local metric learning methods enforce local constraints on the structure of the common multimodal space. Rather than relying on global constraints — like CCA which aims at maximizing the correlation between modalities — these methods adopt a *ranking* point of view, by ensuring that two corresponding elements are close, while two unrelated elements are far. As the goal is to learn a distance d between elements, these approaches are under the umbrella of *metric learning* (Xing et al. 2002).

In a *constrastive* (or *pairwise*) loss, given two elements (v, t) from both modalities, the objective is to (i) minimize their distance $d(v, t)$ if they match ($y = 1$) and (ii) to maximize it if they do not ($y = 1$), by ensuring that their distance is superior to a margin $\gamma > 0$. This is done using a hinge-loss function (Hadsell et al. 2006):

$$\mathcal{L}_{pairwise} = \sum_{y,x_1,x_2} y.d(x_1, x_2) + (1 - y).\lfloor \gamma - d(x_1, x_2) \rfloor_+ \tag{2.12}$$

In a *triplet* loss (illustrated in Figure 2.7), the objective is, given a fixed anchor v, to enforce that the distance $d(v, t^p)$ between v and its corresponding t^p is superior to the distance $d(v, t^n)$ between v and a *negative* element t^n, by a margin γ:

$$\mathcal{L}_{triplet} = \sum_{v,t^p,t^n} \lfloor \gamma + d(v, t^p) - d(v, t^n) \rfloor_+ \tag{2.13}$$

The triplet loss in commonly used in many multimodal tasks. For example, Frome et al. 2013 use it in ZSL (see Section 2.3.1) to learn a projection of an image

vector in a textual semantic space. In Socher et al. 2014a, a triplet loss learns an alignment between images and captions. Carvalho et al. 2018 show, in the retrieval field, that triplet losses outperform pairwise losses.

In the present thesis, we explore an alternative at shared space approaches (see Chapter 6). To do so, we propose novel loss functions aiming at preserving the structure of the visual and textual spaces, without learning an explicit projection between them.

2.1.3.3 Multimodal Language Models

Following the success of large language models such as BERT (Devlin et al. 2019) across a variety of NLP tasks, several research efforts have focused on the design of multimodal versions of such language models to address multimodal tasks, such as VQA. Thus, Multimodal Language Models extend standard NLMs — that are applied only on textual data — to multimodal data. They are all extensions of BERT (which is based on the Transfomer architecture), due to the success of BERT in NLP and the success of attention mechanisms in Computer Vision (K. Xu et al. 2015; H. Xu et al. 2016).

The first attempt in that direction was VideoBERT (C. Sun et al. 2019), a joint *video* and text model, is pre-trained on a huge corpus of YouTube videos, and applied to action classification and video captioning tasks on the YouCook II dataset (L. Zhou et al. 2018). The video is treated as a "visual sentence" (each frame being a "visual word") that is processed by the BERT Transformer, using a special token to signal the beginning of the visual sentence. The model is trained with classic BERT objectives, adapted to the multi-modal setting (word and frame masking), along with a multi-modal alignment task (where the goal is to assess whether a video and a sentence are entailed).

Concerning models jointly treating information from images and text, visual features extracted from the image are used as "visual words", and a `[SEP]` special token is employed to separate textual and visual tokens. In the literature, visual features are object features extracted with a Faster R-CNN (S. Ren et al. 2017) – with the notable exception of Kiela et al. 2019 who used pooling layers from a CNN.

A first body of work exploit *single-stream* Transformers in which visual features are incorporated in a BERT-like Transformer: this is the case for VisualBERT (L. H. Li et al. 2019) — illustrated in Figure 2.8 —, VL-BERT (W. Su et al. 2019), Unicoder-VL (G. Li et al. 2019) and B2T2 (Alberti et al. 2019).

Other works, such as ViLBERT (J. Lu et al. 2019) and LXMERT (Tan et al. 2019) — illustrated in Figure 2.9 — have investigated *two-stream* approaches: these models employ modality-specific encoders built on standard Transformer blocks, which are then fused into a cross-modal encoder.

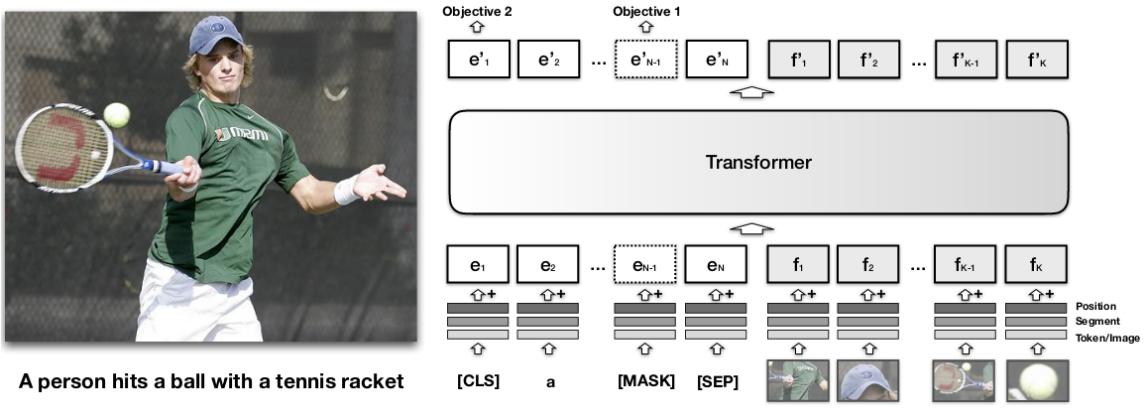

Figure 2.8 – VisualBERT: A Simple and Performant Baseline for Vision and Language. Illustration taken from L. H. Li et al. 2019.

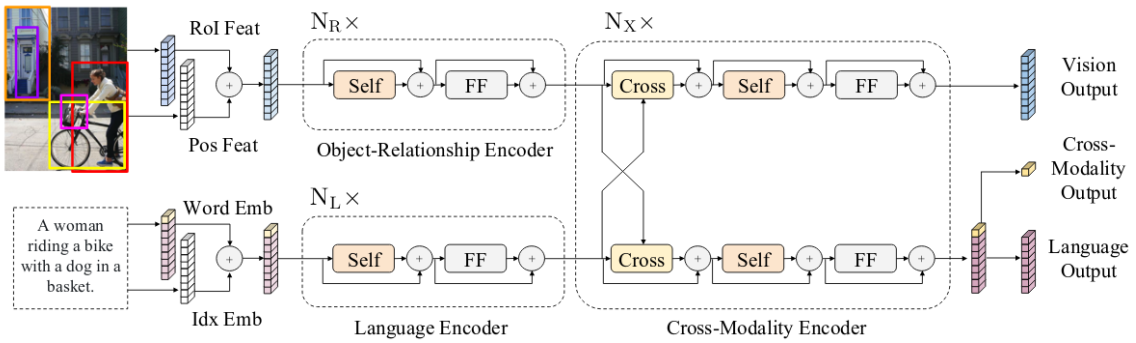

Figure 2.9 – LXMERT: Learning Cross-Modality Encoder Representationsfrom Transformers. Illustration taken from Tan et al. 2019.

In all models, a token embedding encodes the modality, so that visual and textual features can by distinguished by the Transformer encoder.

Among these models, e.g, differences mainly lie in:

1. the pre-training datasets considered — image datasets such as Visual Genome (Krishna et al. 2017), captioning datasets such MS COCO (T. Lin et al. 2014b) or Conceptual Captions (P. Sharma et al. 2018), VQA datasets such as VQA2.0 (Teney et al. 2016), GQA (Hudson et al. 2019), VG-QA (Yuke Zhu et al. 2016)
2. the pre-training tasks — MLM, sentence-to-image matching, masked language feature classification, VQA — and (iii) downstream tasks — visual commonsense reasoning, VQA, image-text retrieval, grounding phrases, etc.

Interestingly, none of the aforementioned models have been used for generation tasks, due to the intrinsic limitation of BERT which is an encoder, and not an encoder-decoder. The construction of a Multimodal Neural Model able to *generate* a sentence using BERT is the purpose of one of our contributions, see Chapter 5.

2.2 NLP aided by Computer Vision (RQ1)

NLP representations, as presented in Section 2.1.1, come from models that have been trained exclusively on textual data. As shown in Section 2.2.1.1, this leads to a lack of common-sense in textual representations. Thus, Computer Vision can be used to enhance NLP representations (Section 2.2.1), but it also can be used to extend current NLP tasks in a multimodal setting, such as Visual Question Generation (VQG) (Section 2.2.2.1), VQA (Section 2.2.2.2), Visual Dialog (Section 2.2.2.3) or MMT (Section 2.2.2.4).

2.2.1 Visual Grounding of Language

2.2.1.1 Motivation

To understand the way language conveys meaning, the traditional approach consists in considering language as a purely symbolic system based on words and syntactic rules (Chomsky 1980; Burgess et al. 1997). However, (Fincher-Kiefer 2001; W. Barsalou 1999) insist on the intuition that language has to be grounded in the real world and perceptual experience. Harnad 1990 illustrates this idea with a thought experiment: imagine that you want to learn Chinese and the only source of information at your disposal is a Chinese dictionary; since you don't know the meaning of any symbol, your learning experience with the dictionary would be an endless and useless go-round.

The importance of real-world grounding is stressed in (Gordon et al. 2013), where an important bias is reported: *the frequency at which objects, relations, or events occur in natural language are significantly different from their real-world frequency*. For example, authors remark that the action of *murdering* is mentioned four times more than an every-day action such as *breathing* in text, as shown in Table 2.7. This can lead to important distortions in artificial common-sense learning. Indeed, common-sense information such as "bananas are yellow" or "the moon is round" are rarely explicitely stated in text, as this type of information is supposed to be known by the (human reader). Similarly, unusual facts such as "the sun rises today" are not mentioned as they are not surprising. However, machines that based their understanding of language by studying vast text corpora might not capture common-sense knowledge. Thus, leveraging visual resources, in addition to textual resources, is a promising way to acquire common-sense knowledge (X. Lin et al. 2015; Yatskar et al. 2016) and cope with the bias between text and reality.

Cognitive psychology and neuroscience works have shown that the human understanding of language is heavily grounded in perception. Indeed, humans, through their senses, have access to perceptual information when learning of

Word	*Teraword*	*Knext*	*Word*	*Teraword*	*Knext*
Spoke	11,577,917	372,042	Hugged	610,040	11,453
Laughed	3,905,519	179,395	Blinked	390,692	21,973
Murdered	2,843,529	16,890	Was late	2,843,529	16,890
Inhaled	984,613	5,617	Exhaled	368,922	31,168
Breathed	725,034	41,215	Was on time	23,997	14

Table 2.7 – Illustration of the Human Reporting Bias, reproduced from Gordon et al. 2013. Count of the number of times that *A person may* $< x >$ in Teraword and Knext text corpora.

words, leading some cognitive scientists to state that the meaning of words is embodied in sensory-motor processing (De Vega et al. 2012). For example, Pulvermüller 2005 show that reading/hearing action verbs such as *kick* or *lick* activates the regions of the brain corresponding to these actions (the leg and the tongue regions, respectively). In Therriault et al. 2009, subjects are presented an image of a pumpkin, and are asked to name the object; when the image is orange, the word *pumpkin* is said rapidly by the subjects, but when the image is grayscale, the answer is slowed down; when the image has the wrong color (e.g, blue), the answer is slowed down even more.

In the following, we present techniques to learn textual representations at two levels of granularity: words in Section 2.2.1.2 and sentences in Section 2.2.1.3.

2.2.1.2 Multimodal word representations

This section explains how to build general-purpose word meaning representations (embeddings) for words, constructed using information from several modalities, in this context: language and images. The intuition behind those models is that textual and visual data have systematic biases with respect to the way they encode information about concepts. Thus, the exploitation of both modalities may lead to more balanced representations.

The non-textual inputs most commonly used to ground language in perception and/or vision are:

- *Feature-norms* (Silberer et al. 2012; Silberer et al. 2013). Feature norms are a list of attributes for an object. They include physical and functional properties associated with the referents of words. The typical datasets for that are the McRae (McRae et al. 2005) dataset (property norms for 500 concrete nouns, with 2,526 properties in total) and the CSLB property norms (semantic properties for 638 concepts) (Devereux et al. 2013).

- *Co-occurrence patterns of words in image tags*: Some works use the indirect grounding of language in images using co-occurrence patterns of words

in image tags and captions. This is the case of the Word2Vec extension model proposed in Felix Hill et al. 2014a where the visual inputs are in fact perceptual features given in the ESP-Game dataset (Ahn et al. 2005) (10,000 images each annotated with a list of lexical concepts that appear in the image, 20,515 distinct tags with an average of 4 tags per image). Felix Hill et al. 2014b and Bruni et al. 2012a use the ESP-Game dataset as well.

- *Natural images*. Bag of Visual Words (BoVW) with SIFT features have been widely used until the raise of representations obtained with CNN. Kiela et al. 2014b shows that CNN features are better suited than BoVW for multimodal semantic evaluation on tasks such as word similarity and relatedness evaluation.

Based on these inputs, a variety of multimodal word representation models have been designed. They can be divided into two groups: joint models and sequential models.

Joint models: Early Fusion Joint models directly learn a joint representation from textual and visual inputs.

In *Bayesian* techniques, the main assumption is that, in documents (e.g., online news), images and surrounding texts have been generated using a shared set of latent variables or topics. Most Bayesian techniques are extensions of Latent Dirichlet Allocation (LDA): the topics are inferred from the joint distribution of textual and visual words, as in Yansong Feng et al. 2010; Roller et al. 2013; Silberer et al. 2013.

Autoencoders have also been used. For example, in Silberer et al. 2014, by concatenating learned representations from two unimodal autoencoders (one for text and one for the visual modality), a multimodal embedding is learned. A semi-supervised criterion is added to perform a classification task based on the representation: this allows to learn representations capable of discriminating between different objects. This model can be used for classification and to infer a modality if it is missing (based on the other one).

Some models propose an extension of the Skip-Gram model (Mikolov et al. 2013b). Felix Hill et al. 2014a base their model on the assumption that frequency of appearance of concrete concepts correlates with the likelihood of "experiencing" it in the world. Thus, perceptual information about a concrete concept is introduced to the model when-ever that concept is encountered in textual modality. Based on external sources, perceptual information is associated with concrete concepts. Concrete words representations are then trained to predict context as in the classical Word2Vec approach and to predict the perceptual features. This amounts to linguistic-context re-weighting.

A. Lazaridou et al. 2015a present the Multi-Modal Skip-Gram model, illustrated in Figure 2.10. It is an extension of the Word2Vec Skip-Gram model and use

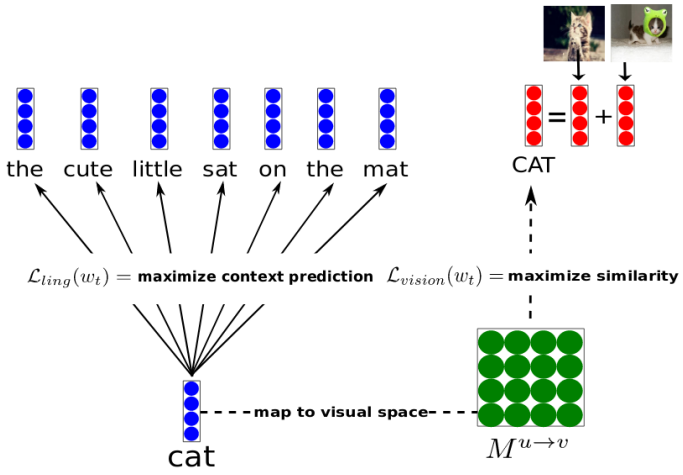

Figure 2.10 – Combining Language and Vision with a Multimodal Skip-gram Model. Illustration taken from A. Lazaridou et al. 2015a.

authentic image analysis as proxy to perceptual information — instead of the feature-norms used in Felix Hill et al. 2014a. The global objective is a linear combination of a Word2vec objective $\mathcal{L}_{ling}$ (Equation 2.2) and a visual objective $\mathcal{L}_{vision}$ which is a ranking triplet loss (Equation 2.13) that brings together the projection of word embeddings to their corresponding visual embeddings.

Sequential Models Sequential models separately construct visual and textual representations and then combine them using different techniques.

A simple way to combine textual and visual representations is to *concatenate* them (Silberer et al. 2013; Bulat et al. 2016) (Section 2.1.3.1). This fusion method is also known as *middle fusion*. In Kiela et al. 2014a, textual representations (learned with Word2Vec skip-gram) are concatenated to visual representations (obtained with the pre-softmax features given by a pre-trained Convolutional Neural Network).

In Collell et al. 2017, a cross-modal projection function f is learned to map word embeddings to their corresponding visual representations — average of 100 CNN representations using images retrieved with Google Images. This approach is illustrated in Figure 2.11. With this model, even abstract words can benefit from visual grounding. The multimodal embedding m_w of a word w is obtained by concatenating the word embedding t_w and its projection $f(t_w)$:

$$m_w = t_w \oplus f(t_w) \tag{2.14}$$

Bruni et al. 2012b and Bruni et al. 2014 consider Singular Value Decomposition (SVD) as a way to fuse modalities. The textual and visual representations are first

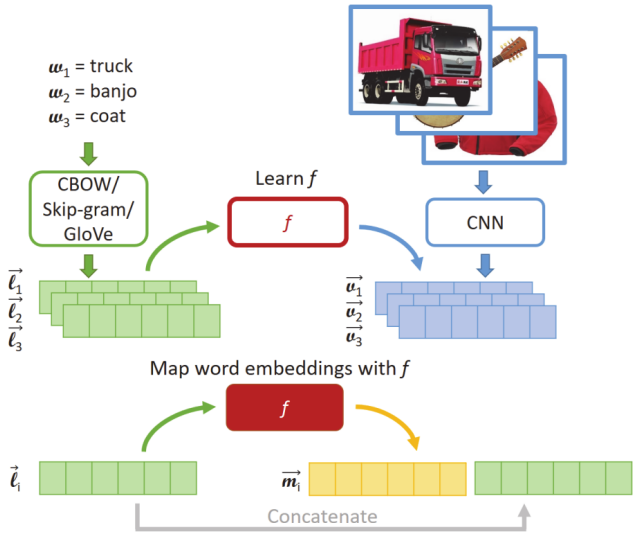

Figure 2.11 – Imagined Visual Representations as Multimodal Embeddings. Illustration taken from Collell et al. 2017.

concatenated and then projected onto a lower dimensionality latent space using SVD.

Silberer et al. 2012, Silberer et al. 2013 and Felix Hill et al. 2014b consider Canonical Correlation Analysis (Section 2.1.3.2) as a way to project textual and visual spaces in a common one. Loeub et al. 2016 propose the Residual CCA method as an improvement: they suggest that important information is also to be found in the dissimilar components of the mono-modal signals. These components are lost in the common space learned by the classical CCA. To use these dissimilar components, they use a residual approach and consider $t_r = t - t'$ and $v_r = v - v'$ where t' and v' are projections of t and v in the common representation space given by the CCA.

Reichart et al. 2013 and Felix Hill et al. 2014b use *weighted gram matrix combination* to fuse linguistic and perceptual information. Given a modality, a weighted gram matrix is constructed. $L_{ij} = S(F_i, F_j).\phi(r_i).\phi(r_j)$ were S is a cosine similarity and ϕ is a quality score of the representations, which reflects the importance of a concept relative to other concepts. Each word representation in the set is thus mapped into a new space of dimension determined by the concept list. There are several advantage to this method: (i) the relative nature of semantics (we generally require models to determine relations between concepts relative to others) is directly encoded since representations are projected onto a space defined by the set of concepts themselves; (ii) dense and fixed size representations are obtained. The final fusion embedding is obtained with a symmetric product of linguistic and perceptual weighted gram matrices.

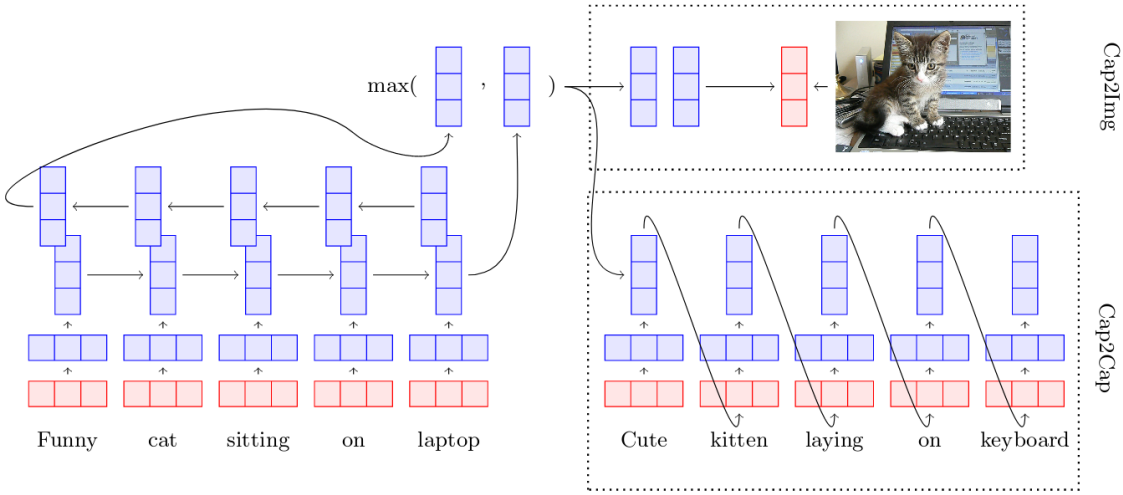

Figure 2.12 – The GroundSent model for learning visually grounded sentence representations. Illustration taken from Kiela et al. 2018.

2.2.1.3 Multimodal sentence representations

While the literature on multimodal word representations is abundant, there are comparably fewer works on the visual grounding of sentences. In this thesis, we present a contribution on this task: this is the purpose of Chapter 6.

Chrupala et al. 2015 propose the IMAGINET model that learns a visually grounded language model. A multi-task objective is used to simultaneously predict the visual representation of a sentence and the next word in the sentence (language model). This model uses the order of the words thanks to the two GRU recurrent networks and learns meaning representations for individual words.

The GroundSent model of Kiela et al. 2018 is close to IMAGINET and additionally hypothesizes that associated captions ground the meaning of a sentence. This model is illustrated in Figure 2.12. In this work, the authors learn a bidirectional LSTM f_θ, parameterized by θ, to encode sentences. Their model is sequential: the multimodal sentence representation is the concatenation of (i) a purely-textual SkipThought vector obtained from textual data, and (ii) grounded sentence vectors obtained with two objectives (that can be combined): Cap2Cap and Cap2Img, trained on a captioning dataset (MS COCO) $\mathcal{D} = (I, S)$, where each image I is associated with its caption S.

Cap2Cap ensures that sentences with a similar visual meaning share a common representation. For two sentences S and S' describing a similar image, Cap2Cap relies on an encoder-decoder framework: the input sentence S is encoded by the sentence encoder f_θ to give the vector $f_\theta(S)$, which is used to condition an encoder that predicts sequentially the words of sentence S'.

Cap2Img ensures that visual semantics are incorporated in sentence representations. The loss function is a max-margin ranking objective, aiming at bringing together the projection of the sentence representations $f_\theta(S)$ to their corresponding image I (represented with the penultimate layer of a pre-trained ResNet) in a common multimodal representation space.

2.2.2 Extensions of NLP tasks using visual information

In this section, we describe various tasks that extend standard NLP tasks by using complementary visual information.

2.2.2.1 Visual Question Generation

The text-based Question Generation task has been largely studied by the NLP community (Rus et al. 2010; Rajpurkar et al. 2016; Q. Zhou et al. 2017; Du et al. 2017; Song et al. 2017; Y. Zhao et al. 2018; Scialom et al. 2019). However, its visual counterpart, VQG, has been comparatively less explored than standard well-known multi-modal tasks such as VQA (H. Xu et al. 2016), Visual Dialog (Das et al. 2017a), or Image Captioning (X. Chen et al. 2015). VQG is thus an extension of a purely-textual task: Textual Generation, which is why we classify it in "NLP aided by CV". However, the "Cross-Modal" description could also be applied to VQG. Indeed, in most VQG papers (Patro et al. 2018a; Patro et al. 2019; Patro et al. 2020), several configurations are considered: (i) the input is a caption of the image — this setting corresponds to the standard Textual Generation task, (ii) the input is the caption *and* the image — this setting corresponds to a standard multimodal extension of a NLP task (such as Multimodal Machine Translation for example), and (iii) the input is the image — this setting corresponds to a cross-modal configuration, where a modality is translated into another modality (except that here the generated text is not a description/caption, but is of a different nature).

From a practical standpoint, the VQG task has several applications: robots or AI assistants could ask questions rooted in multi-modal data (e.g. fusing conversational data with visual information from captors and cameras), in order to refine their interpretation of the situation they are presented with. It could also allow systems relying on knowledge-bases to gain visual common sense and deal with the Human Reporting Bias (Misra et al. 2016), which states that the content of images and text are intrinsically different, since visual common sense is rarely explicitly stated in text.

The VQG task was first introduced by Y. Yang et al. 2015 in their Neural Self Talk model: the goal is to gain knowledge about an image by iteratively generating questions (VQG) and answering them (VQA). The authors tackle the task with a simple RNN conditioned on the image.

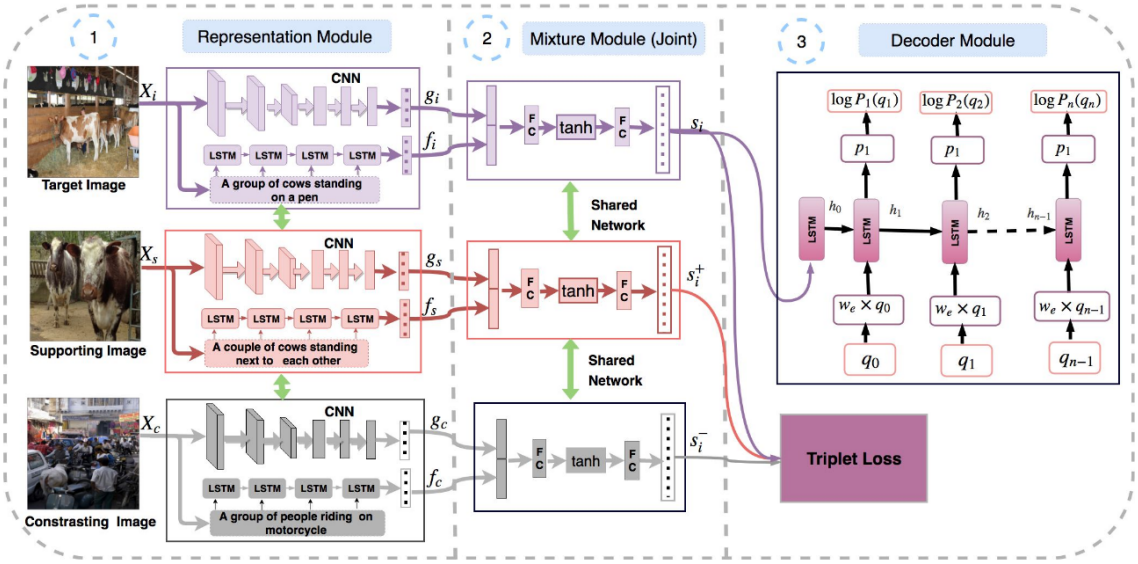

Figure 2.13 – Multimodal Differential Network for Visual Question Generation. Illustration taken from Patro et al. 2018a.

Suitable data for the VQG task can come from standard image datasets on which questions have been manually annotated, such as VQG_{COCO}, VQG_{Flickr}, VQG_{Bing} (Mostafazadeh et al. 2016) , each consisting of 5000 images with 5 questions per image. Alternatively, VQG samples can be derived from VQA datasets, such as $VQA1.0$ (Teney et al. 2016), by "reversing" them (taking images as inputs and questions as outputs).

A variety of approaches have been proposed. Mostafazadeh et al. 2016 use a standard Gated Recurrent Neural Network, *i.e.* a CNN encoder followed by a GRU decoder to generate questions. S. Zhang et al. 2017 aim at generating, for a given image, multiple visually grounded questions of varying types (*what*, *when*, *where*, etc.); similarly, Jain et al. 2017 generate diverse questions using Variational Auto-Encoder (VAE). In Y. Li et al. 2018, VQG is jointly tackled with its dual task (VQA), just as Y. Yang et al. 2015. In (Patro et al. 2018a; Patro et al. 2019), the image (processed by a CNN) and the caption (processed by a LSTM) are combined in a fusion module, followed by a LSTM decoder to generate the question, leading to state-of-the-art results on the VQG task on $VQA1.0$ data. This approach is illustrated in Figure 2.13. More recently, Patro et al. 2020 incorporate multiple cues – place information obtained from PlaceCNN (B. Zhou et al. 2018), caption, tags – and combine them within a deep Bayesian framework where the contribution of each cue is weighted to predict a question, obtaining the current state-of-the-art results on VQG_{COCO}.

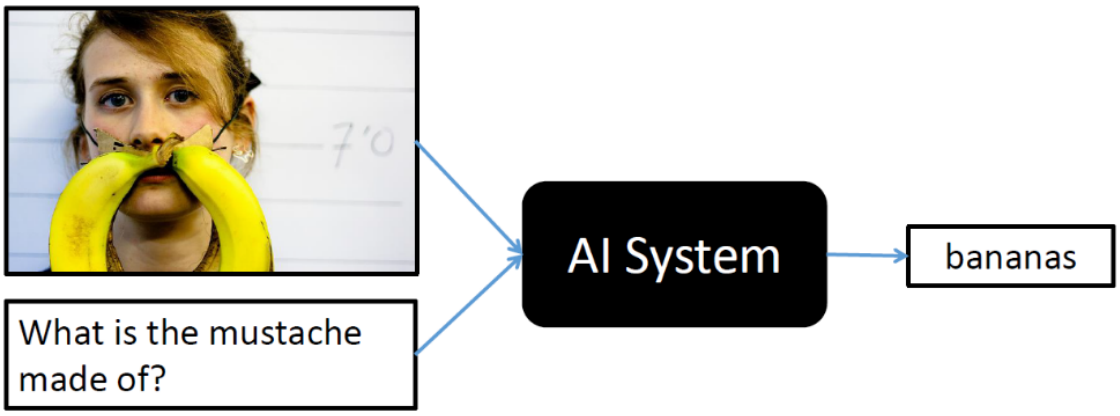

Figure 2.14 – Visual Question Answering. Illustration taken from https://visualqa.org.

2.2.2.2 Visual Question Answering

VQA (Malinowski et al. 2014; Gao et al. 2015; H. Xu et al. 2016; M. Ren et al. 2015; L. Ma et al. 2016; J. Lu et al. 2016; Noh et al. 2016; Fukui et al. 2016; Shih et al. 2016; Patro et al. 2018b) — illustrated in Figure 2.14 — aims at evaluating the visual reasoning capabilities of visual models. Given an image and a question in natural language, the VQA task — proposed in (Malinowski et al. 2014) — aims at determining the right answer among a set of pre-defined answers; it is thus seen as a classification task.

In the first Vanilla VQA model (Antol et al. 2015a), images features, produced with a CNN, and question features, produced with a LSTM, are combined using element-wise operations, later used to determine the correct answer. In the Stacked Attention Network (Z. Yang et al. 2016), an attention mechanism is added using the softmax output of the intermediate question feature, which enables the model to focus on the relevant portion of the image via multiple-step reasoning. Teney et al. 2018 (VQA 2017 challenge winner) first use an object detection model — Faster R-CNN (S. Ren et al. 2017) —in VQA to narrow down visual features and thus produce better attention over the image. Pythia v1.0 (Y. Jiang et al. 2018) (VQA 2018 challenge winner) builds upon Teney et al. 2018 while proposing several improvements: in the model architecture, the learning rate schedule, the image features and data augmentation. In C. Wu et al. 2019, a new module is added in the VQA framework: a Differential Network, whose goal is to refine visual and textual features and reduce observation noise.

The VQA task present several challenges. First, whether learned model have strong visual reasoning capabilities is a central question. In that regard, certain works even target special kinds of reasoning: for example, the images of the Compositional Language and Elementary Visual Reasoning diagnostics (CLEVR) dataset (Johnson et al. 2017) contain 3-D solids of various shapes and colors, and

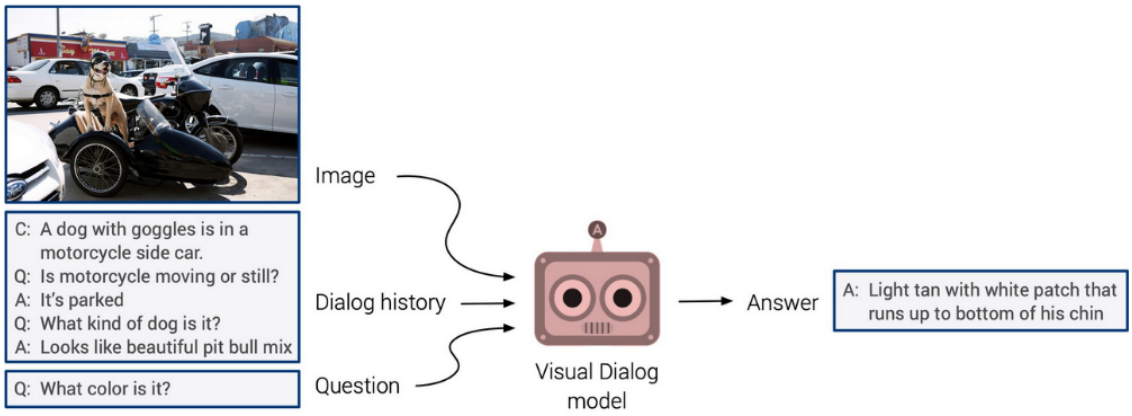

Figure 2.15 – Visual Dialog. Illustration taken from https://visualdialog.org.

the questions involve complex geometrical understanding (e.g, *What size is the cylinder that is left of the brown metal thing that is left of the big sphere?*). Second, the bias toward training data might lead models to ignore the image. Ramakrishnan et al. 2018 indeed show that even purely-textual models, trained only on textual data, manage to get high scores by just focusing on the question. Then, the question of the interpretability of models is important, to understand why models made such predictions; attention modules are a first step toward that goal (Cadène et al. 2019). Finally, multimodal fusion strategies are at the heart of VQA, as both textual input (the question) and visual input (the image) have to be fused to produce an answer, which is also of textual nature.

2.2.2.3 Visual Dialog

Introduced in Das et al. 2017a (who created the VisDial dataset) the Visual Dialog task (Das et al. 2017a; Vries et al. 2017; Strub et al. 2017; Das et al. 2019) requires an AI agent to hold a meaningful dialog with humans in natural language about visual content. In practice, given an image, a dialog history, and a follow-up question about the image, the task is to answer the question. The Visual Dialog task is illustrated in Figure 2.15. The *Visual Object Discovery through Visual Dialog* task, introduced in Vries et al. 2017 along with the GuessWhat?! dataset, is a variation of the Visual Dialog task. The goal is to locate an unknown object in an image by asking a series of natural language question to an agent.

In Visual Dialog, most methods use an encoder-decoder (Sutskever et al. 2014) framework. The encoder model fuses visual and textual information; it can consist of: late-fusion, a hierarchical recurrent network, a memory network (three methods proposed in (Das et al. 2016)), early answer fusion (Jain et al. 2018), history-conditional image attention (J. Lu et al. 2017), and sequential co-attention (Q. Wu et al. 2018). The decoder aims at producing an answer in natural language; it usually consists either of a generative decoder like a RNN (Das et al. 2016) or a

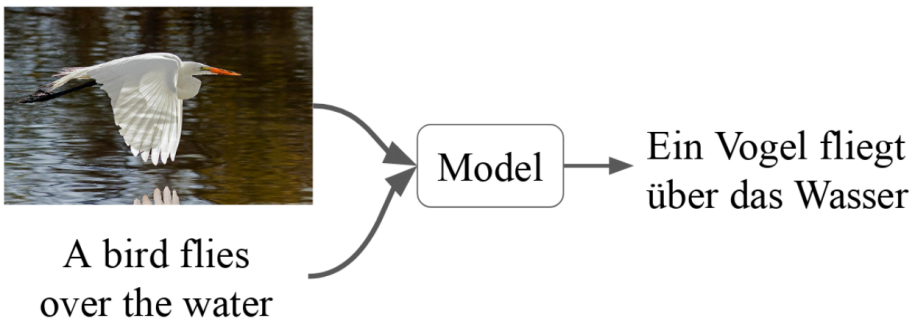

Figure 2.16 – Multimodal Machine Translation. Illustration taken from http://statmt.org/wmt18/multimodal-task.html.

discriminative decoder that ranks answer candidates using a cross-entropy loss (Das et al. 2016) or a ranking-based multi-class loss (J. Lu et al. 2017). Some methods include Reinforcement Learning (Das et al. 2017b; Chattopadhyay et al. 2017) to train two agents to play image guessing games in a collaborative manner. Others use generative models like Generative Adversarial Network (GAN) (Goodfellow et al. 2014) to produce answers that are not distinguishable from human answers (Q. Wu et al. 2018), or VAE (Kingma et al. 2014b) to generate diverse answers (Massiceti et al. 2018).

2.2.2.4 Multi-modal Machine Translation

MMT is an extension of the Machine Translation task, where the input consists of an image and a descriptive caption in language A, and the goal is to predict a translation of the caption in language B. Thus, the image serves as a perceptual reference that can help the model to better understand the original caption, desambiguate some words and thus refine the translation. For example, if the image shows a man running close to a band of water, and the caption is: *A man runs on a river bank*, the word *bank* is polysemous and could refer to the meaning: *financial institution*. However, the image can refine the understanding of *bank*, and the model can translate it by *rive*, in French. The Multimodal Machine Translation task is illustrated in Figure 2.16.

Various approaches have been proposed for MMT: (i) models based on multimodal attention using features extracted by CNNs (Caglayan et al. 2016; Calixto et al. 2016; Libovický et al. 2017; Helcl et al. 2018), (ii) cross-modal interactions with spatially-unaware global features (Calixto et al. 2017; M. Ma et al. 2017) and (iii) the use of regional features extracted with object detection networks (Huang et al. 2016; Grönroos et al. 2018). However, several works have pointed out the fact that the visual modality has not a substantial contribution to the translation performances compared to purely-textual models (Grönroos et al. 2018; Barrault et al. 2018; Lala et al. 2018). Moreover, Elliott 2018 showed that, even when replacing the input image by an unrelated image, the translation performance

does not suffer significant performances. Caglayan et al. 2019 demonstrate that MMT datasets have to be designed so that modalities are complementary rather than redundant: in that case, the visual modality increases the robustness of the machine translation model by mitigating errors in the input sentence.

Even though the contribution of the visual modality is still unclear in the MMT task, some researchers have relied on the assumption that vision could serve as a bridge between languages, as people can always recognize the same objects in the real world, whatever the language. Following this intuition, Nakayama et al. 2017 propose to perform Unsupervised Machine Translation using images as pivot, by learning to project the input sentence into a visual space (using an Image Captioning dataset in language A), and then learning to generate the output sentence from the visual vector (using an Image Captioning dataset in language B). Following the latest developments of Unsupervised Machine Translation (Guillaume Lample et al. 2018a), Y. Su et al. 2019 perform Unsupervised Machine Translation, to help desambiguate the meaning of sentences, using a method close to (Guillaume Lample et al. 2018b).

2.3 Computer Vision aided by NLP (RQ2)

In standard Computer Vision tasks, such as Image Classification (Deng et al. 2009) or object detection task (Uijlings et al. 2013), objects are just defined by a class id, not by their semantics. For example, the output of a standard CNN is a distribution of probability over 1000 pre-defined objects, but it is unable to take into account the semantics of objects to generalize its knowledge to unknown classes (if *dog* is part of the vocabulary but not *puppy*, the model is not able to recognize puppies in images).

This is an important limitation, as the human understanding of the visual world relies on a wide variety of priors that can be found in language. A good example of this fact can be found when a human is confronted to a scene where some object is partially hidden or hard to see. Given the other objects in the context, and using their common-sense knowledge, humans are generally able to narrow down the set of possibilities and determine the correct object — applying such contextual approaches is the purpose of one of our contribution, see Chapter 3.

In this section, we present Computer Vision tasks that benefit from NLP knowledge.

2.3.1 Recognizing visually unknown objects (ZSL)

ZSL is a task that extends a standard Computer Vision task (image classification), to configurations where some classes are unknown to the model. Thus, the model has to take the semantics of object classes into account, using class representations

derived from NLP models. This is why we classify ZSL as a "Computer Vision aided by NLP" task. But it can also be seen as a Cross-Modal Task (see Section 2.4), since it is a Cross-Modal Retrieval task: from a visual input (an image), the goal is to retrieve the corresponding class label, which is of textual nature.

In the present thesis, we present two contributions to the Zero-Shot Learning field. In Chapter 3, we leverage the visual context around objects to refine the predictions of the ZSL model. In Chapter 4, we adapt the CycleGAN model to perform Transductive Zero-Shot Learning.

2.3.1.1 Motivation

Over the last decade, the exponential evolution of computing power, combined with the creation of large-scale image datasets such as ImageNet (Farhadi et al. 2009a), has enabled Convolutional Neural Networks (Lecun et al. 1998) to reach their full power, with recent improvements (Zoph et al. 2018; Real et al. 2019) pushing forward the classification performance every year. However, as noted in (Frome et al. 2013), such models have several drawbacks: they are unable to make predictions that fall outside of the set of training classes, their training requires a large number of examples for each class, and despite outmatching humans on the ImageNet challenge, they are unable to mimic the human capacity to *generalize* prior knowledge to recognize new classes.

2.3.1.2 The ZSL task

To cope with these limitations, *Zero-Shot Learning* approaches (Farhadi et al. 2009b; Mensink et al. 2012; Frome et al. 2013; Fu et al. 2015b; E. Zablocki et al. 2019) have been proposed. In ZSL, two sets of classes are distinguished: the *seen* classes, for which examples are available during training, and the *unseen* classes, for which no labeled images are available. The information learned using seen classes can be generalized to unseen ones by leveraging auxiliary knowledge, which semantically relates seen and unseen classes, e.g. attributes (Ferrari et al. 2007; Parikh et al. 2011; Farhadi et al. 2009b) , or textual embeddings of class labels (Frome et al. 2013). Evaluation is then carried out on the *unseen* classes. The key of ZSL is to use auxiliary knowledge to semantically relate classes from the seen and unseen classes; thus, class labels are embedded in a a common semantic representation space. The ZSL approach is illustrated in Figure 2.17.

The usual procedure in ZSL (Frome et al. 2013) consists in (1) learning a mapping between the visual and the textual space so that images and class labels can be semantically related, (2) performing a nearest neighbor search to find the closest unseen class corresponding to a projected image. Pioneering works focused on hand-crafted attributes for the textual space (Parikh et al. 2011) e.g, 'IsBlack', 'HasClaws'. Since this involves costly and error-prone human labelling, most current works use word vector spaces (Norouzi et al. 2014), such as Word2vec

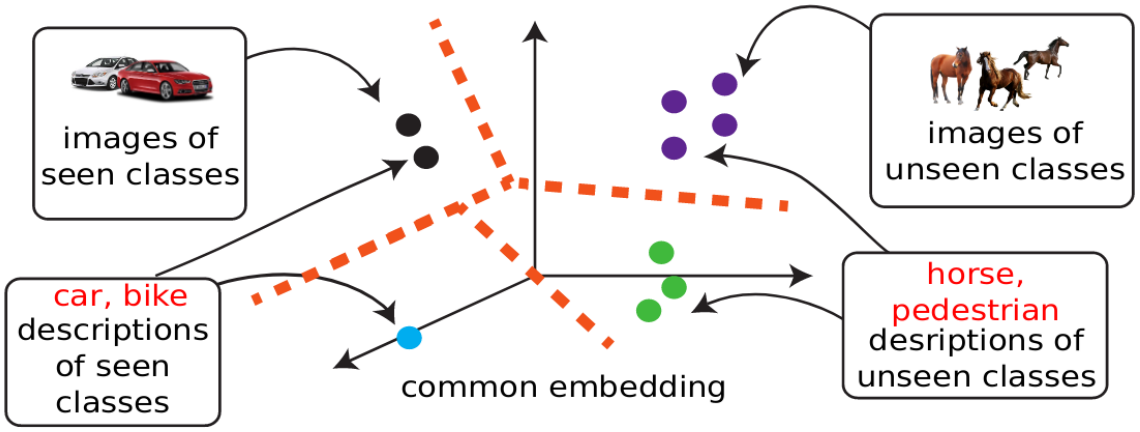

Figure 2.17 – Zero-Shot Learning. Illustration taken from Jurie et al. 2017.

(Mikolov et al. 2013b), which do not suffer from these limitations or Glove (Pennington et al. 2014); concerning images, most ZSL works use CNNs to embed them in a visual embedding space (Fu et al. 2014; Akata et al. 2015; Fu et al. n.d.; Bucher et al. 2016; Romera-Paredes et al. 2015; Z. Zhang et al. 2015; Lampert et al. 2014) by taking the penultimate layer.

Datasets The vast majority of ZSL works (Lampert et al. 2014; Z. Zhang et al. 2015) are evaluated on attribute datasets, namely AWA1 (Farhadi et al. 2009b), AWA2 (Yongqin Xian et al. 2019), CUB (Welinder et al. 2010), SUN (Patterson et al. 2012), aPY (Farhadi et al. 2009b). In these datasets, images are manually annotated given a set of pre-defined attributes, and class vectors are thus derived from these manual annotations; and the total number of classes (both seen and unseen) is relatively small (espectively 50,50,200,717 and 32). The ImageNet dataset is more challenging: it contains 14M images and about 20K unseen classes.

Linear projections The first ZSL approaches learned a linear projection of visual features in the textual space, like DeViSE (Frome et al. 2013). In this model, a max-margin ranking objective is used to learn a cross-modal projection f between an image $\mathcal{V}$ and the semantic representation w_i of its class label i, using the following loss function:

$$\mathcal{L}_{DeViSE} = \sum_{i,\mathcal{V}} \sum_{j} \lfloor \gamma - f(\mathcal{V})^T w_i + f(\mathcal{V})^T w_j \rfloor_+) \tag{2.15}$$

where j is a negative class label sampled uniformly, w_j its representation, and γ is an hyperparameter margin. The DeViSE model is illustrated in Figure 2.18. The DeViSE model serves as the basis for one of our contributions on ZSL, in Chapter 3.

Hybrid Methods Other models, like CONSE (Norouzi et al. 2014) or SYNC (Changpinyo et al. 2016) express image as a mixture of other classes features

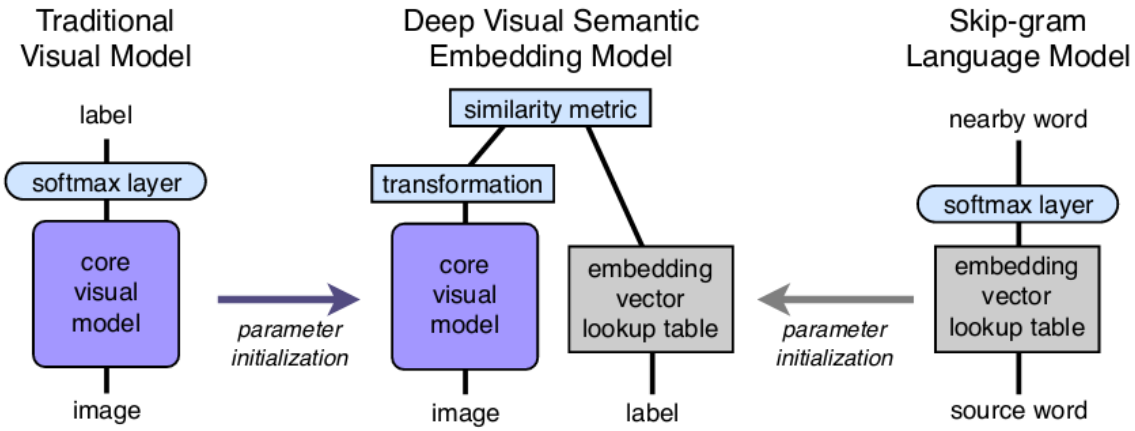

Figure 2.18 – DeViSE: A Deep Visual-Semantic Embedding Model. Illustration taken from Frome et al. 2013.

(*hybrid* models). CONSE (Norouzi et al. 2014) propose to embed images as a convex combination of the word embeddings of the top classes retrieved by a CNN. More precisely, for a given image, let us call $p(i)$ the distribution over the seen classes $i \in [1, |\mathcal{S}_\mathcal{T}|]$ output by the CNN, and T_K the indices of the K most probable classes. The representation v of an image is then

$$v_{\text{CONSE}} = \sum_{i \in T_K} p(i|i \in T_K).w_i \tag{2.16}$$

where w_i is the label of class i and $p(i|i \in T_K) \propto p(i)$. The CONSE model serves as the basis for one of our contributions on ZSL, in Chapter 4.

Going further than CONSE, SYNC (Changpinyo et al. 2016) adopts the point of view of *manifold learning*: classifiers for unseen classes are built by combining classifiers of *phantom* classes, that are embedded both in the semantic space and the model space. In the model space, seen and phantom classes form a weighted bipartite graph.

Non-linear cross-modal functions More recently, non-linear relations between modalities are investigated (Ba et al. 2015; Xian et al. 2016), as in EXEM (Changpinyo et al. 2017) where a kernel-based regressor is learned: it maps semantic representations to visual exemplars while ensuring that the semantic space is clustered efficiently.

Exploiting WordNet information Recent ZSL models tackling the ImageNet ZSL task rely on the exploitation of the WordNet knowledge graph of ImageNet synsets as (X. Wang et al. 2018; Kampffmeyer et al. 2019) using Graph Convolutional Networks (Bruna et al. 2014; Defferrard et al. 2016; Kipf et al. 2017). The

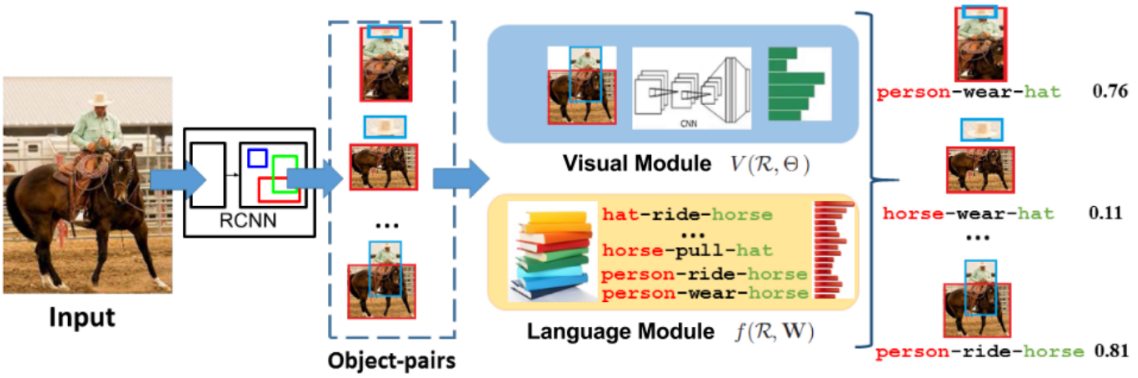

Figure 2.19 – Visual Relationship Detection with Language Priors. Illustration taken from C. Lu et al. 2016a.

complementary information of the hierarchy of ImageNet classes enable these models to reach state-of-the-art performances.

2.3.2 Visual Relationship Detection

Visual Relationship Detection (VRD) (M. A. Sadeghi et al. 2011; C. Lu et al. 2016b; F. Sadeghi et al. 2015; Jung et al. 2019; R. Yu et al. 2017) aims at detecting visual relations in images, in the form of triplets *subject, predicate, object* (e.g, *man, holding, baseball bat*) where *subject* and *object* are elements of the image and *predicate* explains their relationship.

As noted in Jung et al. 2019, VRD faces several challenges. First, many triplets have very rarely (or never) been seen in the training data; thus, understanding separately the meanings of objects and predicates is paramount. Second, an object may take various appearances depending on the triplet: e,g. the object *chicken* in both triplets *man, eating, chicken* and *chicken, eating, corn* has very different appearance. Finally, models are penalized when the predicted answer is not the ground-truth answer, even when both are semantically close (e.g, *under* vs *below* or *close to* vs *next to*).

To tackle these limitations, language knowledge is used in VRD works, in order to distill external textual priors in the visual model. For example, in (C. Lu et al. 2016b), the predicates are projected in a representation space where they are embedded to reflect their semantics; by combining object embeddings with predicate embeddings, even unseen triplets can be given non-null probabilities by the model, and the semantics of the whole triplet is taken into account. In (R. Yu et al. 2017), linguistic statistics extracted from the VRD dtraining dataset and Wikipedia are extracted to estimate the conditional probability of a predicate given a *subject,object* pair; this knowledge is then distilled in the model.

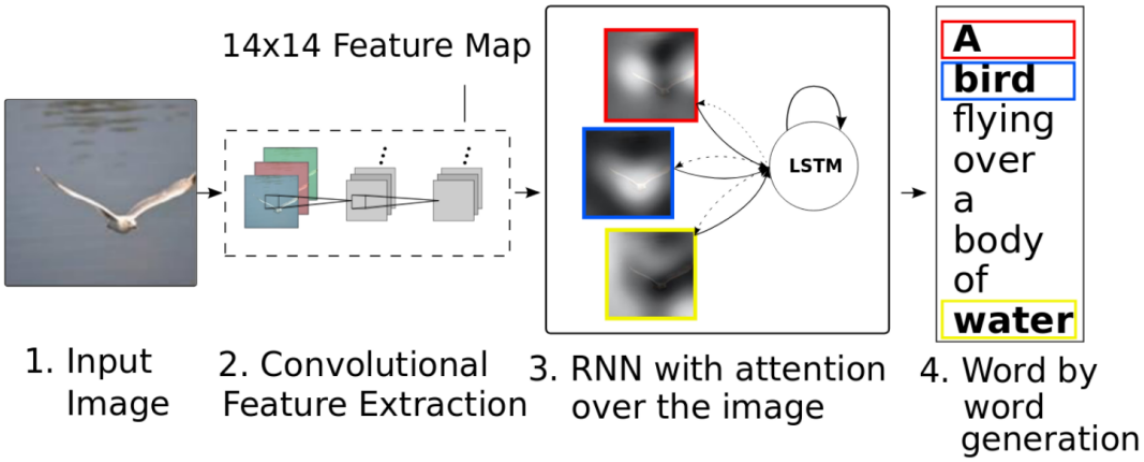

Figure 2.20 – Image Captioning. Illustration taken from K. Xu et al. 2015.

2.4 Cross-Modal Tasks (RQ3)

A *cross-modal task* is a task where a modality is somewhat translated into another. The image-to-text configuration corresponds to the Image Captioning task (Section 2.4.1), and the text-to-image configuration to the Text-to-Image Synthesis task (Section 2.4.2). In the Cross-Modal Retrieval task (Section 2.4.4), both configurations are made possible, except that the goal is not to *generate* an element from the other modality, but to *retrieve* the semantically closest element in a pre-defined set.

2.4.1 Image Captioning

Image Captioning (Socher et al. 2014a; Vinyals et al. 2015b; Karpathy et al. 2017; K. Xu et al. 2015; Fang et al. 2015; X. Chen et al. 2015; Johnson et al. 2016; Yan et al. 2016) aims at generating a description in natural language given an image; thus, it evaluates global scene understanding capabilities of models. The challenges tackled by the Image Captioning task involve: the recognition of objects in the image, understanding their relations (spatial organization, actions, movements), selecting information worth to be mentioned — an important aspect stressed in the Human Reporting Bias (Gordon et al. 2013) — and describing it in fluent natural language. The Image Captioning task is illustrated in Figure 2.20.

Most methods adopt an encoder-decoder framework (Sutskever et al. 2014), in which a CNN encodes an image into a vector in a latent space, and a RNN decodes it to sequentially generate a caption. Attention mechanisms have been used to refine this strategy (K. Xu et al. 2015; Engilberge et al. 2018), so that the model can focus on the relevant parts of the image while generating the caption. As Image Captioning is a text generation task, it is evaluated using standard metrics

like BiLingual Evaluation Understudy (BLEU) (Papineni et al. 2002) and Metric for Evaluation of Translation with Explicit ORdering (METEOR) (Banerjee et al. 2005). Thus, some works (Ranzato et al. 2016; Z. Ren et al. 2017) use Reinforcement Learning to optimize these metrics, which are non-differentiable.

The Image Captioning task is at the origin of some variants. For example, in the Dense Captioning task (Johnson et al. 2016), the goal is to generate a sentence describing a region within the image, rather than the whole image. More recently, the Unsupervised Image Captioning task has been considered (Yang Feng et al. 2019), where the alignment between images and captions is unknown during training. To do so, a sentence corpus teaches the model to generate plausible sentences via adversarial training, while the knowledge of a pre-trained visual concepts detector is distilled into the model via reinforcement learning; moreover, a cycle-consistency objective ensures that generated captions are semantically consistent with the image.

The Image Captioning task suffers from some problems. First, biases have been observed toward the training set: models tend to use contextual cues rather than focusing on the appearance of objects in the image, in particular for genders (Hendricks et al. 2018). Then, evaluation metrics like BLEU and METEOR have intrinsic limitations, pointed out in Novikova et al. 2017, as they do not correlate well with human judgments. Thus, Image Captioning might not be sufficient to evaluate visual reasoning and scene understanding capabilities.

2.4.2 Text-to-Image Synthesis

Text-to-Image Synthesis (Reed et al. 2016b; H. Zhang et al. 2017; Gorti et al. 2018) is the inverse task of Image Captioning: starting from a sentence (e.g, *this is a flower with round purple upward facing petals*), the goal is to generate an image that illustrates that sentence. This task was proposed following the recent developments of GAN (Goodfellow et al. 2014) — that have shown substantial results in image generation (Radford et al. 2016) and image-to-image translation (J. Zhu et al. 2017) — and conditional GANs (Mirza et al. 2014), that learn to approximate the distribution of data by conditioning on an input.

Most Text-to-Image Synthesis works use conditional GANs to generate images conditioned on a textual input (Reed et al. 2016b; H. Zhang et al. 2017; Dash et al. 2017; H. Zhang et al. 2019) — textual descriptions are in general encoded using SkipThought vectors (Kiros et al. 2015) or Char-CNN-RNN embeddings (Reed et al. 2016a). Reed et al. 2016b are the first to condition GANs using textual descriptions instead of class labels. H. Zhang et al. 2019 introduce a two-stage method to generate images. In Stage 1, a GAN G1 conditioned on textual description produces a low-resolution 64*64 image. In Stage 2, the input is the image generated by Stage 1, and another GAN G2 produces a high-resolution 256*256 realistic image. However, as noted in Gorti et al. 2018, generated images don't always reflect

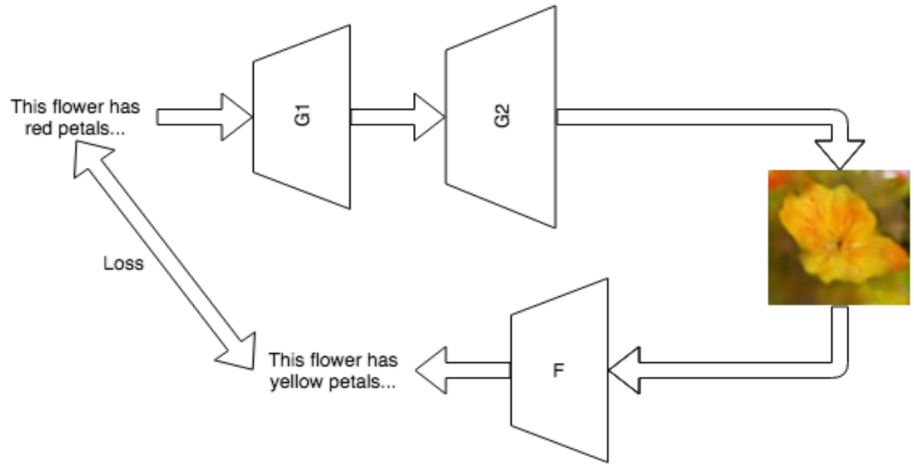

Figure 2.21 – Text-to-Image-to-Text Translation using Cycle Consistent Adversarial Networks. Illustration taken from Gorti et al. 2018. G1, G2 are the stage 1 and stage 2 generator. F is the caption generator.

the meaning of the sentence given as input with the aforementioned models. Following the success of CycleGAN (J. Zhu et al. 2017), that has been largely adopted in uni-modal settings (Y. Lu et al. 2017; Almahairi et al. 2019; J. Zhao et al. 2019), Gorti et al. 2018 propose to use a cycle-consistency loss to ensure a closer correspondence between images and sentences. To do so, they use a captioning network F that, by taking as input the generated image, generates a describing caption, which is then compared to the initial sentence. The latter approach is illustrated in Figure 2.21.

Due to the high difficulty of the task, most papers tackling Text-to-Image Synthesis use datasets that have a limited semantic range, for example the Oxford VGG 102 Flower Dataset (Nilsback et al. 2008) (only flower pictures) or the CUB dataset (Welinder et al. 2010) (only bird images). The Text-to-Image Synthesis is thus more challenging than Image Captioning, as (i) task-wise, generating meaningful images is more difficult than generating correct sentences in natural language, and (ii) evaluation-wise, the set of images corresponding to a given sentence is extremely large.

2.4.3 Grounding phrases in Images

Grounding textual elements in visual data (Karpathy et al. 2017; Kong et al. 2014; Plummer et al. 2017; R. Hu et al. 2016) aims at localizing a word/phrase/sentence/paragraph in an image by determining a a relevant bounding box. Thus, it is close to the object detection task (Uijlings et al. 2013; Girshick et al. 2014; S. Ren et al. 2017), which is a traditional Computer Vision task, except that in this task textual semantics have to be incorporated to understand and localize complex queries such

as: *a small boy* or *a cat eating a mouse*. It should not be confused with the Visual Grounding of Language task, described in Section 2.2.1, which aims at enhancing textual representations using visual information.

The general procedure (R. Hu et al. 2016; Mao et al. 2016) consists in generating candidate locations for objects, and computing the similarity scores between the image regions and the textual query. Thus, it is close to the image-to-sentence matching task (see Section 2.4.4). This task is usually performed using the Flickr30K Entities dataset (Plummer et al. 2015): an augmentation of the Flickr30K dataset (Young et al. 2014), with bounding boxes for all noun phrases present in textual descriptions. Along with the new dataset, Plummer et al. 2015 propose a simple baseline: a CCA model to linearly project image and textual features by maximizing their correlation. L. Wang et al. 2016 introduce a Deep Structure-Preserving Embedding for image-sentence matching that they apply to phrase grounding. R. Hu et al. 2016 propose a Spatial Context Recurrent ConvNet, in which a caption generation model is used to score the phrase on a set of proposed boxes. The GroundR model (Rohrbach et al. 2016) adds a reconstruction loss, that ensures that the retrived bounding box can predict the initial textual phrase, thus penalizing a wrong location choice.

2.4.4 Cross-Modal Retrieval

Cross-Modal Retrieval aims at bridging the visual and textual modalities: either a text is given as input, and the goal is to retrieve the semantically closest image (Text-to-Image), or an image is given as input, and the goal is to retrieve the semantically closest text (Image-to-Text). Due to the explosion of all kind of multimodal content on the Internet (images, video, texts), Cross-Modal Retrieval presents important real-life applications (K. Wang et al. 2016). While first methods used statistical tools, such as Canonical Correlation Analysis (Hardoon et al. 2004) or kernel-based methods (Akaho 2006; Weiran Wang et al. 2016), deep learning approaches have been used to learn the common cross-modal representation space (Andrew et al. 2013; Weiran Wang et al. 2015; Peng et al. 2016; Wei et al. 2017; Peng et al. 2018; Qi et al. 2018; Gu et al. 2018) following the success of Deep Learning (LeCun et al. 2015). Cross-Modal Retrieval is usually performed by representing images and texts in a common semantic space, in which a distance between them can be computed.

A variety of methods have been designed for Cross-Modal Retrieval: they all rely on aligned data between text and images. There are (i) unsupervised approaches (Andrew et al. 2013; F. Feng et al. 2014; Weiran Wang et al. 2015), (ii) pairwise approaches (D. Zhai et al. 2012; X. Zhai et al. 2013; J. Wang et al. 2015) and (iii) supervised approaches (K. Wang et al. 2016; A. Sharma et al. 2012).

In unsupervised methods, the only information that is used to build common data representation are co-occurrence patterns between text and images; for

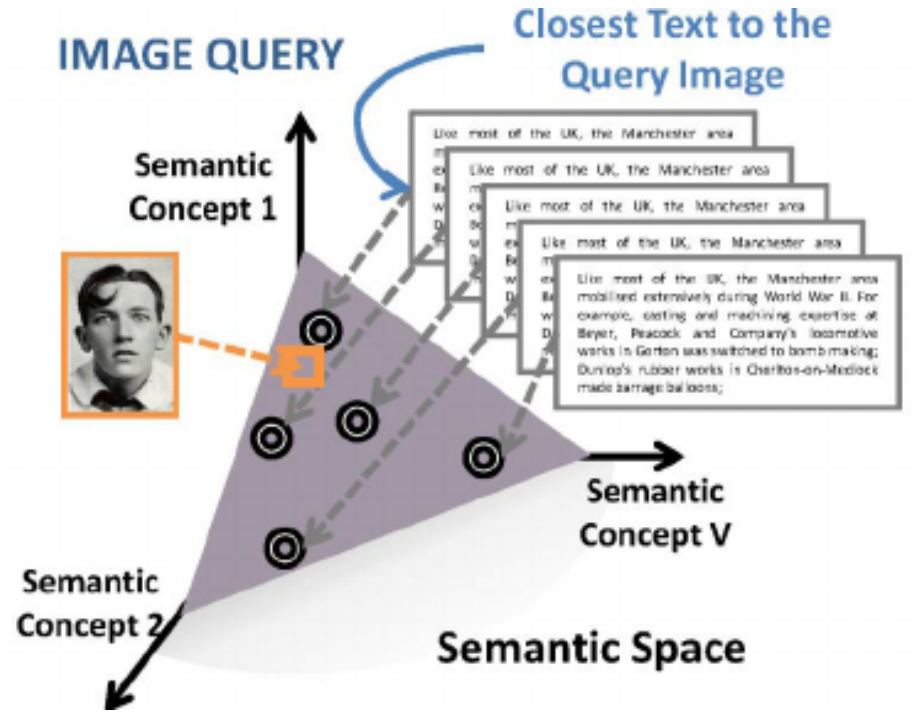

Figure 2.22 – Cross-Modal retrieval using a common semantic space. Illustration taken from Pereira et al. 2014.

example, CCA (see Section 2.1.3.2) and its derivatives Deep CCA (DCCA) (Andrew et al. 2013), Correspondence Auto-encoder (Corr-AE) (F. Feng et al. 2014).

In pairwise approaches, text/image pairs are used to learn a metric able to efficiently compare samples from both modalities; for example, Multi-view Metric Learning with Global consistency and Local smoothness (MVML-GL) method (D. Zhai et al. 2012), the Joint Graph Regularised Heterogeneous Metric Learning (JGRHML) method (X. Zhai et al. 2013) and the Modality-Specific Deep Structure (MSDS) method (J. Wang et al. 2015).

In supervised approaches, the goal is to learn discriminative representations, able to separate classes in the common representation space, by exploiting label information. For example, Wei Wang et al. 2016 present a Multi-modal Deep Neural Network, where images are embedded with a CNN and texts with a NLM, learned with a combination of intra-modal losses (capturing semantic relationships within each modality) or inter-modal losses (capturing semantic relationships across modalities). Some works also use adversarial losses as a complement of supervised losses, to enforce that distributions of both modalities are consistent in the common representation space. Adversarial learning can be performed either as a refinement once a standard triplet max-margin loss has been learned (R. Liu et al. 2019), or since the beginning of training (L. He et al. 2017; B. Wang et al. 2017).

To reduce the storage cost of Cross-Modal Retrieval models, *binary-valued* approaches have been proposed by using binary hash codes. This is done via Cross-Modal Hashing (Kumar et al. 2011; Ding et al. 2014; D. Zhang et al. 2014; Zijia Lin et al. 2015), a method in which data points from one modality are mapped into a Hamming space of binary codes where the similarity in the original space is preserved. Q. Jiang et al. 2017 extend Cross-Modal Hashing by using deep features instead of hand-crafted ones.

2.5 Positioning

In Chapter 3, we propose a contribution to ZSL (Section 2.3.1). We are interested in determining what visual information is encoded in word embeddings (**RQ2**). Standard ZSL assume that textual representations encode information about the visual appearance of objects: this prior linguistic knowledge is used to recognize objects that are unknown to the model. Our goal is to show that these textual representations also encode information about the visual context of objects (i.e. knowledge about which objects co-occur in images) or their visual frequence. To do so, we introduce a new task: *context-aware zero-shot learning*, where the goal is to determine the class of an object (unknown to the model) delimited by a bounding box in the image, while taking into account its *visual context* (**??**).

In Chapter 4, we propose a contribution to Transductive Zero-Shot Learning (T-ZSL), a setting of ZSL where images and class labels of unseen classes are available during training, but the correspondence between them is unknown. In this contribution, we are interested in studying how multimodal tasks can benefit from visual and textual data when supervision is weak (**RQ1**,**RQ2**,**RQ3**) — in addition to T-ZSL, we also tackle Cross-Modal Retrieval and Visual Grounding of Language in settings where text/image is weak, or even non-existent. Current T-ZSL models generally rely on methods that cluster the space of unseen classes, and thus these approaches only work when the number of unseen classes is relatively low. We tackle this limitation by proposing an approach based on the CycleGAN model (J. Zhu et al. 2017), where the distribution of unseen classes is aligned to the distribution of their corresponding images with adversarial learning.

In Chapter 5, we present a contribution on Visual Question Generation (Section 2.2.2.1). We explore whether BERT representations can generalize their knowledge to a multimodal task (**RQ2**), in that case: VQG. Thus, we extend the BERT model to a generation framework, that we call *BERTgen*, and incorporate visual information within the model as if it were of textual nature.

In Chapter 6, we tackle Visual Grounding of Language (**RQ1**) and propose to learn multimodal sentence representations (Section 2.2.1.3). We explore an alternative to shared space approaches (Section 2.1.3.2), as we argue that a shared

space over-constrains the learned space in the case of sentences. To do so, we introduce two objectives aimed at preserving the structure of both spaces in an intermediate space.

CHAPTER 3

LEVERAGING VISUAL KNOWLEDGE WITHIN LANGUAGE FOR COMPUTER VISION

Contents

Chapter abstract

Zero-Shot Learning (ZSL) aims at classifying unlabeled objects by leveraging auxiliary knowledge, such as semantic representations. A limitation of previous approaches is that only intrinsic properties of objects, e.g. their visual appearance, are taken into account while their context, e.g. the surrounding objects in the image, is ignored. Following the intuitive principle that objects tend to be found in certain contexts but not others, we propose a new approach, context-aware ZSL, that leverages semantic representations in a

new way to model the conditional likelihood of an object to appear in a given context. Finally, through extensive experiments conducted on Visual Genome, we show that contextual information can substantially improve the standard ZSL approaches and is robust to unbalanced classes.

The work in this Chapter has led to the publication of a conference paper:

- *Eloi Zablocki, Patrick Bordes, Laure Soulier, Benjamin Piwowarski, and Patrick Gallinari (2019). "Context-Aware Zero-Shot Learning for Object Recognition". In: ICML 2019.*

3.1 Introduction

Traditional Computer Vision models, such as Convolutional Neural Networks (CNNs) (Lecun et al. 1998), are designed to classify images into a set of predefined classes. Their performances have kept improving in the last decade, namely on object recognition benchmarks such as ImageNet (Deng et al. 2009), where state-of-the-art models (Zoph et al. 2018; Real et al. 2019) have outmatched humans. However, training such models requires hundreds of manually-labeled instances for each class, which is a tedious and costly acquisition process. Moreover, these models cannot replicate humans' capacity to generalize and to recognize objects they have never seen before. As a response to these limitations, Zero-Shot Learning (ZSL) has emerged as an important research field in the last decade (Farhadi et al. 2009b; Mensink et al. 2012; Fu et al. 2015b; Kodirov et al. 2017). Zero-Shot Learning has been covered in the Background Chapter of this thesis (Section 2.3.1). In the object recognition field, ZSL aims at identifying an object class for which no supervised data is available, by using knowledge acquired from another disjoint set of classes, for which corresponding visual instances are provided. In the literature, these sets of classes are respectively called *target* and *source* domains — terms borrowed from the transfer learning community. Generalization from the source to the target domain is achieved using auxiliary knowledge that semantically relates classes of both domains, e.g. attributes or textual representations of the class labels.

Previous ZSL approaches only focus on intrinsic properties of objects, e.g. their visual appearance, by the means of handcrafted features — e.g. shape, texture, or color — (Lampert et al. 2014) or distributed representations learned from text corpora (Akata et al. 2016; Y. Long et al. 2017). The underlying hypothesis is that the identification of entities of the target domain is made possible thanks to the implicit *principle of compositionality*, a.k.a. Frege's principle (Pelletier 2001) — an object is formed by the composition of its attributes and characteristics — and the fact that other entities of the source domain share the same attributes. For example, if textual resources state that an apple is round and that it can be red or green, this knowledge can be used to identify apples in images because these

characteristics ('round', 'red') could be shared by classes of the source domain (e.g, 'round' like a ball, 'red' like a strawberry...).

We believe that visual context, i.e. the other entities surrounding an object, also explains human's ability to recognize an object that has never been seen before. This assumption relies on the fact that scenes are *compositional* in the sense that they are formed by the composition of objects they contain. Some works in Computer Vision have exploited visual context to refine the predictions of classification (Mensink et al. 2014) or detection (Bell et al. 2016) models. The use of contextual information in Computer Vision has been detailed in the Background Chapter, **??**. To the best of our knowledge, context has not been exploited in ZSL because, for obvious reasons, it is impossible to directly estimate the likelihood of a context for objects from the target domain — from visual data only. However, textual resources can be used to provide insights on the possible visual context in which an object is expected to appear. To illustrate this, knowing from language that an apple is likely to be found hanging on a tree or in the hand of someone eating it, can be very helpful to identify apples in images.

In this paper, our goal is to leverage visual context as an additional source of knowledge for ZSL, by exploiting the distributed word representations (Mikolov et al. 2013b) of the object class labels. More precisely, we adopt a probabilistic framework in which the probability to recognize a given object is split into three components:

- a *visual component* based on its visual appearance (which can be derived from previous ZSL models),
- a *contextual component* exploiting its visual context,
- a *prior component*, which estimates the frequency of objects in the dataset.

As a complementary contribution, we show that separating prior information in a dedicated component, along with simple yet effective sampling strategies, leads to a more interpretable model, able to deal with imbalanced datasets. Finally, as traditional ZSL datasets lack contextual information, we design a new dedicated setup based on the richly annotated Visual Genome dataset (Krishna et al. 2017). We conduct extensive experiments to thoroughly study the impact of contextual information.

3.2 Chapter Questions

This Chapter is linked to the following Research Questions, as defined in the Introduction of the present thesis:

- **RQ2** (*Can language help to refine visual understanding ?*): The main assumption in ZSL is that language representations contain information about the visual

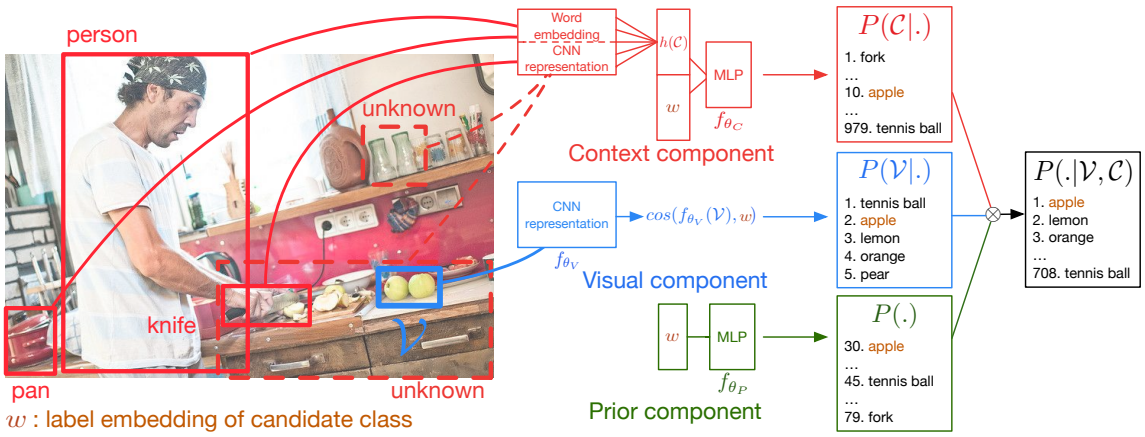

Figure 3.1 – The goal is to find the class (in the target domain) of the object contained within the blue image region $\mathcal{V}$. Its context is formed of labeled objects from the source domain (red plain boxes) and of unlabeled object from the target domain (red dashed boxes).

appearance of objects. In this Chapter, we challenge this assumption by proposing a new one: *language representations also contain information about the visual context of objects in images, as well as on the frequence of objects in images*. This assumption leads us to a new task (context-aware ZSL) and a new model that we evaluate extensively.

- **RQ3** (*Can modalities be translated into one another?*): ZSL is a Cross-Modal task, as an image is translated into a class label. In this Chapter, we are interested in showing that *the quality of this translation depends on the criterion that we choose*. Indeed, our model is modular, with three separate and interpretable modules, that are specialized on different criterions: visual appearance, visual context and frequence. We propose a method to weigh the contributions of each module and propose an optimal image-to-text translation.

As a result, we derive two Chapter Questions (CQ) that we strive to answer throughout this Chapter:

- **CQ1**: What visual information about objects is encoded in textual representations ?
- **CQ2**: How can cross-modal models be more interpretable ?

3.3 Context-aware Zero-Shot Learning

In the present section, we formalize the context-aware Zero-Shot Learning setting.

Let $\mathcal{O}$ be the set of all object classes, divided in classes from the *source domain* $\mathcal{S}$ and classes from the *target domain* $\mathcal{T}$. The goal of our approach — *context-aware* ZSL — is to determine the class $i \in \mathcal{T}$ of an object contained in an image I, given its visual appearance $\mathcal{V}$ and its visual context $\mathcal{C}$. The image I is annotated with bounding boxes, each containing an object. Given the zone $\mathcal{V}$ containing the object of interest, the context $\mathcal{C}$ consists of the surrounding objects in the image. Their classes can either belong to the source domain ($\mathcal{C} \cap \mathcal{S}$) or to the target domain ($\mathcal{C} \cap \mathcal{T}$). Note that the class of an object of $\mathcal{C} \cap \mathcal{T}$ is not accessible in ZSL, only its visual appearance is.

3.3.1 Model overview

We tackle this task by modeling the conditional probability $P(i|\mathcal{V}, \mathcal{C})$ of a class i given both the visual appearance $\mathcal{V}$ and the visual context $\mathcal{C}$ of the object of interest. Given the absence of data in the target domain, we need to limit the complexity of the model, for generalizability's purpose. Accordingly, we suppose that $\mathcal{V}$ and $\mathcal{C}$ are conditionally independent given the class i — we show in the experiments (Section 3.5) that this hypothesis is acceptable. This hypothesis leads to the following expression:

$$P(i|\mathcal{V}, \mathcal{C}) \propto P(\mathcal{V}|i)P(\mathcal{C}|i)P(i) \tag{3.1}$$

where each conditional probability expresses the probability of either the visual appearance $\mathcal{V}$ or the context $\mathcal{C}$ given class i, and $P(i)$ denotes the prior distribution of the dataset. Each term of this equation is modeled separately.

The intuition behind our approach is illustrated in Figure 3.1, where the blue box contains the object of interest. Here, the class is *apple*, which belongs to the target domain $\mathcal{T}$. The visual component, which focuses on the zone $\mathcal{V}$, recognizes a *tennis ball* due to its yellow and round appearance; *apple* is ranked second. The prior component indicates that *apple* is slightly more frequent than *tennis ball*, but the frequency discrepancy may not be high enough to change the prediction of the visual component. In that case, the context component is discriminant: it ranks objects that are likely to be found in a kitchen, and reveals that an *apple* is far more likely to be found than a *tennis ball* in this context.

Precisely modeling $P(\mathcal{C}|.)$, $P(\mathcal{V}|.)$ and $P(.)$ is challenging due to the ZSL setting. Indeed, these distributions cannot be computed for classes of the target domain because of the absence of corresponding training data. Thus, to transfer the knowledge acquired from the source domain to the target domain, we use a common semantic space, namely *Word2Vec* (Mikolov et al. 2013b), where source and target class labels are embedded as vectors of $\mathbb{R}^d$, with d the dimension of the space. It is worth noting that we propose to separately learn the prior class distribution $P(.)$ with a ranking loss (in Section 3.3.3). This allows dealing with

imbalanced datasets, in contrast to ZSL models like DeViSE (Frome et al. 2013). This intuition is experimentally validated in Section 3.5.2.

3.3.2 Description of the model's components

Due to both the ZSL setting and the variety of possible context and/or visual appearance of objects, it is not possible to estimate directly the different probabilities of Equation 3.1. Hence, in what follows, we estimate quantities related to $P(\mathcal{C}|.)$, $P(\mathcal{V}|.)$ and $P(.)$ using parametric energy functions (LeCun et al. 2006). These quantities are learned separately, as described in Section 3.3.3. Finally, we explain how we combine them to produce the global probability $P(.|\mathcal{C}, \mathcal{V})$ in Section 3.3.4.

Visual component The visual component models $P(\mathcal{V}|i)$ by computing the compatibility between the visual appearance $\mathcal{V}$ of the object of interest, and the semantic representation w_i of the class i.

Following previous ZSL works based on cross-modal projections (Frome et al. 2013; Bansal et al. 2018), we introduce f_{θ_V}, a parametric function mapping an image to the semantic space:

$$f_{\theta_V}(\mathcal{V}) = W_V.\texttt{CNN}(\mathcal{V}) + b_V \in \mathbb{R}^d \tag{3.2}$$

where $\texttt{CNN}(\mathcal{V})$ is a vector in $\mathbb{R}^{d_{\text{visual}}}$, output by a pretrained CNN truncated at the penultimate layer, W_V is a projection matrix ($\in \mathbb{R}^{d \times d_{\text{visual}}}$) and b_V a bias vector — in our experiments, $d_{\text{visual}} = 2048$. The probability that the image region $\mathcal{V}$ corresponds to the class i is set to be proportional to the cosine similarity between the projection $f_{\theta_V}(\mathcal{V})$ of $\mathcal{V}$ and the semantic representation $w_i \in \mathbb{R}^d$ of i:

$$\log P(\mathcal{V}|i;\theta_V) \propto \cos(f_{\theta_V}(\mathcal{V}), w_i) := \log \widetilde{P}_{visual} \tag{3.3}$$

Context component The context component models $P(\mathcal{C}|i)$ by computing a compatibility score between the visual context $\mathcal{C}$, and the semantic representation w_i of class i. More precisely, the conditional probability is written:

$$\begin{aligned}\log P(\mathcal{C}|i;\theta_C) \propto f_{\theta_C}(\mathcal{C}, w_i) &= f_{\theta_C^1}\left(h_{\theta_C^2}(\mathcal{C}) \oplus w_i\right) \\ &:= \log \widetilde{P}_{context}\end{aligned} \tag{3.4}$$

where $h_{\theta_C^2}(\mathcal{C}) \in \mathbb{R}^d$ is a vector representing the context, $\theta_C = \{\theta_C^1; \theta_C^2\}$ are parameters to learn, and $\oplus$ is the concatenation operator. To take non-linear and high-order interactions between $h_{\theta_C^2}(\mathcal{C})$ and w_i into account, $f_{\theta_C^1}$ is modeled by a 2-layer Perceptron. We found that concatenating $h_{\theta_C^2}(\mathcal{C})$ with w_i leads to better results than a cosine similarity, as done in Equation 3.3 for the visual component.

To specify the modeling of $h_{\theta^2_C}(\mathcal{C})$, we propose various *context models* depending on which context objects are considered and how they are represented. Specifically, a context model is characterized by (a) the domain of context objects that are considered (i.e. source $\mathcal{S}$ or target $\mathcal{T}$) and (b) the way these objects are represented, either by a textual representation of their class label or by a visual representation of their image regions. Accordingly, we distinguish:

• The *low-level* (L) approach that computes a representation from the image region $\mathcal{V}_k$ of a context object. This produces the following context models:

$$S_L = \{W_C \texttt{CNN}(\mathcal{V}_k) + b_C | k \in \mathcal{C} \cap \mathcal{S}\}$$
$$T_L = \{W_C \texttt{CNN}(\mathcal{V}_k) + b_C | k \in \mathcal{C} \cap \mathcal{T}\}$$

• The *high-level* (H) approach which considers semantic representations w_k of the class labels k of the context objects (only available for entities of the source domain). This produces context models:

$$S_H = \{w_k | k \in \mathcal{C} \cap \mathcal{S}\} \text{ and } T_H = \{w_k | k \in \mathcal{C} \cap \mathcal{T}\}$$

Note that T_H is not defined in the zero-shot setting, since class labels of objects from the target domain are unknown; yet it is used to define Oracle models (Section 3.4.3).

These four basic sets of vectors can further be combined in various ways to form new context models (for instance: $S_L \cup T_L, S_H \cup S_L, S_H \cup S_L \cup T_L$, etc.). At last, $h_{\theta^2_C}$ averages the representations of these vectors to build a global context representation. For example, $h_{\theta^2_C}(\mathcal{C}_{S_H \cup T_L})$ equals:

$$\frac{1}{|\mathcal{C}_\mathcal{S}| + |\mathcal{C}_\mathcal{T}|}\Big[\sum_{(i,\mathcal{V}_i)\in\mathcal{C}_\mathcal{S}} w_i + \sum_{(j,\mathcal{V}_j)\in\mathcal{C}_\mathcal{T}} \big(W_C.\texttt{CNN}(\mathcal{V}_j) + b_C\big)\Big]$$

where $|\cdot|$ denotes the cardinality of a set of vectors.

Context Model	Word emb. known obj.	CNN rep. known obj.	CNN rep. unknown obj.
K_t	✓		
K_v		✓	
U_v			✓
$K_v + U_v$		✓	✓
$K_t + U_v$	✓		✓
$K_{v+t} + U_v$	✓	✓	✓

Table 3.1 – Context models

Prior component The goal of the prior component is to assess whether an entity is frequent or not in images. We estimate $P(i)$ from the semantic representation w_i of class i:

$$\log P(i;\theta_P) \propto f_{\theta_P}(w_i) := \log \widetilde{P}_{prior} \tag{3.5}$$

where f_{θ_P} is a 2-layer Perceptron that outputs a scalar.

3.3.3 Learning

In this section, we explain how we learn the energy functions f_{θ_C}, f_{θ_V} and f_{θ_P}. Each component (resp. context, visual, prior) of our model is assigned a training objective (resp. $\mathcal{L}_C$, $\mathcal{L}_V$, $\mathcal{L}_P$). As the components are independent by design, they are learned separately. This allows for a better generalization in the target domain, as shown experimentally (Section 3.5.2). Besides, ensuring that some configurations are more likely than others motivates us to model each objective by a max-margin ranking loss, in which a positive configuration is assigned a lower energy than a negative one, following the *learning to rank* paradigm (Weston et al. 2011). Unlike previous works (Frome et al. 2013), which are generally based on balanced datasets such as ImageNet and thus are not concerned with prior information, we want to avoid any bias coming from the imbalance of the dataset in $\mathcal{L}_C$ and $\mathcal{L}_V$, and learn the prior separately with $\mathcal{L}_P$. In other terms, the visual (resp. context) component should focus exclusively on the visual appearance (resp. visual context) of objects. This is done with a careful sampling strategy of the negative examples within the ranking objectives, that we detail in the following. To the best of our knowledge, such a discussion relative to prior modeling in learning objectives — which is, in our view, paramount in imbalanced datasets such as Visual Genome — has not been done in previous research.

Positive examples are sampled among entities of the source domain from the data distribution $P^\star$: they consist in a single object for $\mathcal{L}_P$, an object/box pair for $\mathcal{L}_V$, an object/context pair for $\mathcal{L}_C$. To sample negative examples j from the source domain, we distinguish two ways:

(1) For the prior objective $\mathcal{L}_P$, negative object classes are sampled from the *uniform* distribution U:

$$\mathcal{L}_P = \mathop{\mathbb{E}}_{i\sim P^\star} \mathop{\mathbb{E}}_{j\sim U} \left\lfloor \gamma_P - f_{\theta_P}(w_i) + f_{\theta_P}(w_j) \right\rfloor_+ \tag{3.6}$$

Noting $\Delta_{ji} := f_{\theta_P}(w_j) - f_{\theta_P}(w_i)$, the contribution of two given objects i and j to this objective is:

$$P^\star(i) \left\lfloor \gamma_P + \Delta_{ji} \right\rfloor_+ + P^\star(j) \left\lfloor \gamma_P - \Delta_{ji} \right\rfloor_+$$

If $P^\star(i) > P^\star(j)$, i.e. when object class i is more frequent than object class j, this term is minimized when $\Delta_{ji} = -\gamma_P$, i.e. $f_{\theta_P}(w_i) = f_{\theta_P}(w_j) + \gamma_P > f_{\theta_P}(w_j)$. Thus, $\widetilde{P}_{prior}(.;\theta_P)$ captures prior information, as it learns to rank objects based on their frequency.

(2) For the visual and context objectives, negative object classes are sampled from the prior distribution $P^\star(.)$:

$$\mathcal{L}_V = \mathop{\mathbb{E}}_{i,\mathcal{V}\sim P^\star} \mathop{\mathbb{E}}_{j\sim P^\star} \left\lfloor \gamma_V - f_{\theta_V}(\mathcal{V})^\top w_i + f_{\theta_V}(\mathcal{V})^\top w_j \right\rfloor_+ \tag{3.7}$$

$$\mathcal{L}_C = \mathop{\mathbb{E}}_{i,\mathcal{C}\sim P^\star} \mathop{\mathbb{E}}_{j\sim P^\star} \left\lfloor \gamma_C - f_{\theta_C}(\mathcal{C}, w_i) + f_{\theta_C}(\mathcal{C}, w_j) \right\rfloor_+ \tag{3.8}$$

Similarly, the contribution of two given objects i, j and a context $\mathcal{C}$ to the objective $\mathcal{L}_C$ is:

$$P^\star(i)P^\star(j)\Big[P^\star(\mathcal{C}|i)\left\lfloor \gamma_V + f_{\theta_C}(\mathcal{C}, w_j) - f_{\theta_C}(\mathcal{C}, w_i)\right\rfloor_+ + P^\star(\mathcal{C}|j)\left\lfloor \gamma_V + f_{\theta_C}(\mathcal{C}, w_i) - f_{\theta_C}(\mathcal{C}, w_j)\right\rfloor_+\Big]$$

Minimizing this term does not depend on the relative order between $P^\star(i)$ and $P^\star(j)$; thus, $\widetilde{P}_{context}(\mathcal{C}|.;\theta_C)$ does not take prior information into account. Moreover, $P^\star(\mathcal{C}|i) > P^\star(\mathcal{C}|j)$ implies that $f_{\theta_C}(\mathcal{C}, w_i) > f_{\theta_C}(\mathcal{C}, w_j)$.

The alternative, as done in DeViSE (Frome et al. 2013), is to sample negative classes uniformly in the source domain in the objective $\mathcal{L}_V$. Thus, if the prior is uniform, DeViSE directly models $P(.|\mathcal{V})$; otherwise, $\mathcal{L}_V$ cannot be analyzed straightforwardly. Besides, the contributions of visual and prior information are mixed. However, we show that learning the prior separately and imposing the context (resp. visual) component to exclusively focus on contextual (resp. visual) information is more efficient (Section 3.5.2).

3.3.4 Inference

In this section, we detail the inference process. The goal is to combine the predictions of the individual components of the model to form the global probability distribution $P(.|\mathcal{V},\mathcal{C})$. In Section 3.3.3, we detailed how to learn the functions f_{θ_C}, f_{θ_V} and f_{θ_P}, from which $\log \widetilde{P}_{context}$, $\log \widetilde{P}_{visual}$ and $\log \widetilde{P}_{prior}$ are deduced respectively. However, the normalization constants in Equation 3.3, Equation 3.4 and Equation 3.5, which depend on the object class i in the general case, are unknown. As a simplifying hypothesis, we suppose that these normalization constants are scalars that we respectively note α_C, α_V and α_P. This leads to:

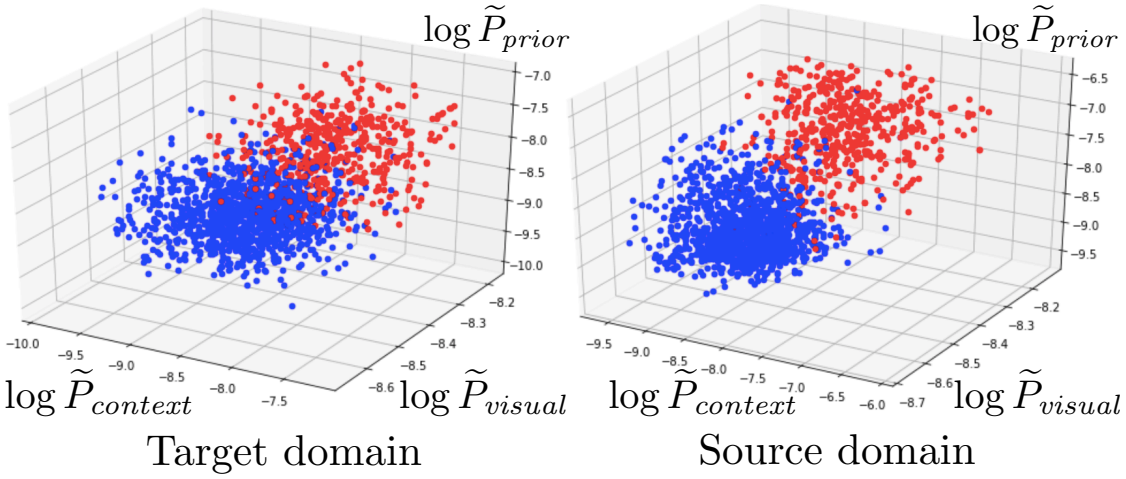

Figure 3.2 – 3D visualization of the unnormalized log-probabilities of each component ($N = 500$). Context model $S_L \cup S_H \cup T_L$.

$$P(.|\mathcal{V},\mathcal{C}) \propto \underbrace{(\widetilde{P}_{context})^{\alpha_C}}_{P(\mathcal{C}|.)} \cdot \underbrace{(\widetilde{P}_{visual})^{\alpha_V}}_{P(\mathcal{V}|.)} \cdot \underbrace{(\widetilde{P}_{prior})^{\alpha_P}}_{P(.)} \tag{3.9}$$

To see whether this hypothesis is reasonable, we did some *post-hoc* analysis of one of our model, and plotted in Figure 3.2 the values $\log \widetilde{P}_{visual}$, $\log \widetilde{P}_{context}$ and $\log \widetilde{P}_{prior}$ for positive (red points) and negative (blue points) configurations $(i,\mathcal{V},\mathcal{C})$ of the test set of Visual Genome. We observe that positive and negative triplets are well separated, which empirically validates our initial hypothesis.

Hyper-parameters α_C, α_V and α_P are selected on the validation set to compute $P(.|\mathcal{C},\mathcal{V})$. To build models that do not use a visual/contextual component, we simply select a subset of the probabilities and their respective hyperparameters. For example:

$$P(.|\mathcal{C}) \propto (\widetilde{P}_{context})^{\alpha_C} (\widetilde{P}_{prior})^{\alpha_P} \tag{3.10}$$

3.4 Experimental protocol

3.4.1 Data

To measure the role of context in ZSL, a dataset that presents annotated objects within a rich visual context is required. However, traditional ZSL datasets, such as AwA (Farhadi et al. 2009b), CUB-200 (Wah et al. 2011) or LAD (B. Zhao et al. 2018), are made of images that contain a unique object each, with no or very

little surrounding visual context. We rather use Visual Genome (Krishna et al. 2017), a large-scale image dataset (108K images) annotated at a fine-grained level (3.8M object instances), covering various concepts (105K unique object names). This dataset is of particular interest for our work, as objects have richly annotated contexts (31 object instances per image on average). In order to shape the data to our task, we randomly split the set of images of Visual Genome into train/validation/test sets (70%/10%/20% of the total size).

To build the set $\mathcal{O}$ of all objects classes, we select classes which appear at least 10 times in Visual Genome and have an available *Word2vec* representation. $\mathcal{O}$ contains 4842 object classes; it amounts to 3.4M object instances in the dataset. This dataset is highly imbalanced as 10% of most represented classes amount to 84% of object instances. We define the *level of supervision* p_{sup} as the ratio of the size of the source domain over the total number of objects:

$$p_{\text{sup}} = |\mathcal{S}|/|\mathcal{O}| \tag{3.11}$$

For a given p_{sup} ratio, the source $\mathcal{S}$ and target $\mathcal{T}$ domains are built by randomly splitting $\mathcal{O}$ accordingly. Every object is annotated with a bounding box and we use this supervision in our model for entities of both source and target domains. To facilitate future work on context-aware ZSL, we publicly release data splits and annotations [1].

3.4.2 Evaluation methodology and metrics

We adopt the conventional setting for ZSL, which implies entities to be retrieved only among the target domain $\mathcal{T}$. Besides, we also evaluate the performance of the model to retrieve entities of the source domain $\mathcal{S}$ (with models tuned on the target domain).

The model's prediction takes the form of a list of n classes, sorted by decreasing probability; the rank of the correct class in that list is noted r. Depending on the setting, n equals $|\mathcal{T}|$ or $|\mathcal{S}|$. We define the First Relevant (FR) metric with:

$$\text{FR} = \frac{2}{n-1}(r-1) \tag{3.12}$$

To further evaluate the performance over the whole test set, the Mean First Relevant (MFR) metric is used (Fuhr 2017). It is computed by taking the mean value of First Relevant (FR) scores obtained on each image of the test set. Note that the factor $\frac{2}{n-1}$ rescales the metric such that the MFR score of a random baseline is 100%, while the MFR of a perfect model would be 0%. The MFR metric has the

1. https://data.lip6.fr/context_aware_zsl/

advantage to be interval-scale-based, unlike more traditional Recall@k metrics or Mean Reciprocal Rank (MRR) metrics (Ferrante et al. 2017), and thus can be averaged; this allows for meaningful comparison with a varying p_{sup}.

3.4.3 Scenarios and Baselines

Model scenarios Model scenarios depend on the information that is used in the probabilistic setting: $\varnothing$, $\mathcal{C}$, $\mathcal{V}$ or both $\mathcal{C}$ and $\mathcal{V}$. When contextual information is involved, a context model $\star$ is specified to represent $\mathcal{C}$, which we note $\mathcal{C}_\star$. The different context models are $\star \in \{S_H, S_L, T_L, S_L \cup T_L, S_H \cup T_L, S_L \cup S_H \cup T_L\}$. For clarity's sake, we note our model M. For example, $\text{M}(\mathcal{C}_{S_H \cup T_L}, \mathcal{V})$ models the probability $P(\mathcal{C}_{S_H \cup T_L}|.)P(\mathcal{V}|.)P(.)$ as explained in Section 3.3.4, $\text{M}(\mathcal{V})$ models $P(\mathcal{V}|.)P(.)$, and $\text{M}(\varnothing)$ models $P(.)$.

Oracles To evaluate upper-limit performances for our models, we define Oracle baselines where classes of target objects are used, which is not allowed in the zero-shot setting. Note that every Oracle leverages visual information. We consider the following Oracle models:

- *True Prior:* This Oracle uses, for its prior component, the true prior distribution $P^\star(i) \propto \frac{\#i}{M}$ computed for all objects of both source and target domains on the full dataset, where $\#i$ is the number of instances of the i-th class in images and M is the total number of images.

$$P_{TruePrior}(i|\mathcal{C}, \mathcal{V}) = P(\mathcal{V}|i)P^\star(i) \propto P(\mathcal{V}|i).\frac{\#i}{M} \tag{3.13}$$

- *Visual Bayes:* This Oracle uses $P^\star(.)$ for its prior component as well. Its context component uses co-occurrence statistics between objects computed on the full dataset: $P^{\text{im}}(\mathcal{C}|i) = \prod_{c \in \mathcal{C}} P_{\text{co-oc}}(c|i)$ where $P_{\text{co-oc}}(c|i) = \frac{\#(c,i)}{\#c\#i}$ is the probability that objects c and i co-occur in images, with $\#(c, i)$ the number of co-occurrences of c and i.

$$P_{VisualBayes}(i|\mathcal{C}, \mathcal{V}) = P^{\text{im}}(\mathcal{C}|i)P(\mathcal{V}|i)P^\star(i) \propto \prod_{c \in \mathcal{C}} \frac{\#(c,i)}{\#c\#i}.P(\mathcal{V}|i).\frac{\#i}{M} \tag{3.14}$$

- *Textual Bayes:* Inspired by (S. Bengio et al. 2013), this Oracle is similar to Visual Bayes, except that its prior $P^{\text{text}}(.)$ and context component $P^{\text{text}}(.|\mathcal{C})$ are based on textual co-occurrences instead of image co-occurrences: $P_{\text{co-oc}}(c|i)$ is computed by counting co-occurrences of words c and i in windows of size 8 in the Wikipedia dataset, and $P^{\text{text}}(i)$ is computed by summing the number of instances of the i-th class divided by the total size of Wikipedia.

$$P_{TextualBayes}(i|\mathcal{C},\mathcal{V}) = P^{\text{text}}(\mathcal{C}|i)P(\mathcal{V}|i)P^{\text{text}}(i) \tag{3.15}$$

- *Semantic representations for all objects:* $\text{M}(\mathcal{C}_{S_H \cup T_H}, \mathcal{V})$ uses word embeddings of both source and target objects. It is an oracle model because, in ZSL, we do not have access to the class labels of target objects.

Baselines We consider the following baselines:

- $\text{M}(\mathcal{C} \oplus \mathcal{V})$: To study the validity of the hypothesis about the conditional independence of $\mathcal{C}$ and $\mathcal{V}$, we introduce a baseline where we directly model $P(\mathcal{C},\mathcal{V}|.)P(.)$. To do so, we replace, in the expression of $\mathcal{L}_V$ (Equation 3.7), $f_{\theta_V}(\mathcal{V})$ by the concatenation of $h(\mathcal{C})$ and $f_{\theta_V}(\mathcal{V})$ projected in $\mathbb{R}^d$ with a 2-layer Perceptron.

- $\text{DeViSE}(\mathcal{V})$: To evaluate the impact of our Bayesian model (Equation 3.1) and our sampling strategy (Section 3.3.3), we compare against DeViSE (Frome et al. 2013). $\text{DeViSE}(\mathcal{V})$ is different from $\text{M}(\mathcal{V})$ because negative examples in $\mathcal{L}_V$ are uniformly sampled, and the prior $P(.)$ is not learned.

- $\text{DeViSE}(\mathcal{C} \oplus \mathcal{V})$: similarly to $\text{M}(\mathcal{C} \oplus \mathcal{V})$, we define a baseline that does not rely on the conditional independence of $\mathcal{C}$ and $\mathcal{V}$, using the same sampling strategy as DeViSE. This is done by replacing, in the expression of the visual loss $\mathcal{L}_\mathcal{V}$ (Equation 3.7), $f_{\theta_V}(\mathcal{V})$ by $f_{\theta_V}(\mathcal{V}) \oplus h(\mathcal{C})$, and by projecting this vector in $\mathbb{R}^d$ using a 2-layer Multi-Layer Perceptron (MLP).

- $\text{M}(\mathcal{C}_I, \mathcal{V})$: To understand the importance of context supervision, i.e. annotations of context objects (boxes and classes), we design a baseline where no context annotations are used. The context is the whole image without the zone $\mathcal{V}$ of the object, which is masked out. The associated context model is $\star = I$ with $h(\mathcal{C}_\mathcal{I}) = g_{\theta_I}(I \setminus \mathcal{V})$; g_{θ_I} is a parametric function to be learned. This baseline is inspired from (Torralba et al. 2010), where global image features are used to refine the prediction of an image model.

3.4.4 Implementation details

For each objective $\mathcal{L}_C$, $\mathcal{L}_V$ and $\mathcal{L}_P$, at each iteration of the learning algorithm, 5 negative entities are sampled per positive example. Word representations are vectors of $\mathbb{R}^{300}$, learned with the Skip-Gram algorithm (Mikolov et al. 2013b) on Wikipedia. Image regions are cropped, rescaled to (299×299), and fed to CNN, an Inception-v3 CNN (Szegedy et al. 2016a), whose weights are kept fixed during training. This model is pretrained on ImageNet (Farhadi et al. 2009a). As a result, every ImageNet class that belongs to the total set of objects $\mathcal{O}$ was included in the source domain $\mathcal{S}$. Models are trained with Adam (Kingma et al.

		Target domain $\mathcal{T}$			Source domain $\mathcal{S}$		
	p_{sup}	10%	50%	90%	10%	50%	90%
Model	Probability						
Random	$\mathcal{U}$	*100*	*100*	*100*	*100*	*100*	*100*
$\text{M}(\varnothing)$	$P(.)$	38.6	23.7	13.8	12.0	10.6	11.2
$\text{M}(\mathcal{V})$	$P(\mathcal{V}\|.)P(.)$	20.5	10.7	6.0	1.5	2.6	3.6
$\text{M}(\mathcal{C}_{S_H})$	$P(\mathcal{C}\|.)P(.)$	28.7	14.4	9.1	4.2	4.3	4.4
$\text{M}(\mathcal{C}_{S_H}, \mathcal{V})$	$P(\mathcal{C}\|.)P(\mathcal{V}\|.)P(.)$	**18.1**	**9.0**	**5.2**	**1.1**	**1.9**	**2.4**
$\delta_{\mathcal{C}}$ (%)		*11.6*	*16.4*	*12.1*	*23.7*	*27.3*	*31.5*

Table 3.2 – **Evaluation of various information sources, with varying levels of supervision.** MFR scores in %. $\delta_{\mathcal{C}}$ is the relative improvement (in %) of $\text{M}(\mathcal{C}_{S_H}, \mathcal{V})$ over $\text{M}(\mathcal{V})$. Entities are retrieved only among entities of the domain at hand.

2014a) and regularized with a L2-penalty; the weight of this penalty decreases when the level of supervision increases, as the model is less prone to overfitting. All hyper-parameters are cross-validated on classes of the target domain, on the validation set.

3.5 Results

3.5.1 The importance of context

In this section, we evaluate the contribution of contextual information, with varying levels of supervision p_{sup}. We fix a simple context model (the model S_H, which uses high-level information of source classes) and report MFR results with $p_{\text{sup}} = 10, 50, 90\%$ in Table 3.2 for every combination of information sources: $\varnothing$, $\mathcal{V}$, $\mathcal{C}$ and $(\mathcal{C}, \mathcal{V})$ — we observe similar trends for the other context models.

Results highlight that:

- Contextual knowledge acquired from the source domain can be transferred to the target domain, as $\text{M}(\mathcal{C}_{S_H})$ significantly outperforms the *Random* baseline.
- As expected, it is not as useful as visual information: $\text{M}(\mathcal{V}) \overset{\text{MFR}}{<} \text{M}(\mathcal{C}_{S_H})$, where $\overset{\text{MFR}}{<}$ means lower MFR scores, i.e. better performances.
- However, Table 3.2 demonstrates that contextual and visual information are complementary: $\text{M}(\mathcal{C}_{S_H}, \mathcal{V})$ outperforms both $\text{M}(\mathcal{C}_{S_H})$ and $\text{M}(\mathcal{V})$.
- Interestingly, as the learned prior model $\text{M}(\varnothing)$ is also able to generalize, we show that visual frequency can somehow be learned from textual semantics,

Model	Probability	p_{sup}	Target domain $\mathcal{T}$ 10%	50%	90%	Source domain $\mathcal{S}$ 10%	50%	90%
Random	$\mathcal{U}$		*100*	*100*	*100*	*100*	*100*	*100*
M($\varnothing$)	$P(.)$		39.6	26.3	16.9	6.6	8.68	10.9
M($\mathcal{V}$)	$P(\mathcal{V}\|.)P(.)$		21.0	11.8	6.9	0.9	2.3	3.5
M($\mathcal{C}_{S_H}$)	$P(\mathcal{C}\|.)P(.)$		28.6	15.0	10.7	3.5	3.9	4.4
M($\mathcal{C}_{S_H}, \mathcal{V}$)	$P(\mathcal{C}\|.)P(\mathcal{V}\|.)P(.)$		18.2	9.4	6.0	0.8	1.8	2.4
$\delta_{\mathcal{C}}$ (%)			*13.4*	*20.2*	*13.4*	*13.8*	*24.4*	*31.5*

Table 3.3 – **MFR scores in the generalized ZSL setting**. Entities are retrieved among every possible entities (from both the source and target domain)

which extends previous work where word embeddings were shown to be a good predictor of textual frequency (Schakel et al. 2015).

When p_{sup} increases, we observe that all models are better at retrieving objects of the target domain (i.e. MFR decreases), which is intuitive because models are trained on more data and thus generalize better to recognize entities from the target domain.

Besides, when p_{sup} increases, the context is also more abundant. This explains:

- the decreasing MFR values for model M($\mathcal{C}_{S_H}$) on $\mathcal{T}$,
- the increasing relative improvement δ_C of M($\mathcal{C}_{S_H}, \mathcal{V}$) over M($\mathcal{V}$) on $\mathcal{S}$.

However, on the target domain, we note that δ_C does not monotonously increase with p_{sup}. A possible explanation is that the visual component improves faster than the context component, so the relative contribution brought by context to the final model M($\mathcal{C}_{S_H}, \mathcal{V}$) decreases after $p_{\text{sup}} = 50\%$. Since the highest relative improvement δ_C (in $\mathcal{T}$) is attained with $p_{\text{sup}} = 50\%$, we fix the standard level of supervision $p_{\text{sup}} = 50\%$ in the rest of the experiments; this amounts to 2421 classes in both source and target domains.

In Table 3.3, we report results obtained when both source and target object classes exist in the retrieval space: this setting amounts to *generalized zero-shot learning*. The aforementioned observations still hold true in the generalized setting. Indeed, due to the nature of the retrieval metric, for a given model, MFR score are extremely close whether retrieval is performed in the source domain or in both the source and target domain.

	Model	Probability	$\mathcal{T}$	$\mathcal{S}$
Oracles	*Textual Bayes*	$P^{\text{text}}(\mathcal{C}\|.)P(\mathcal{V}\|.)P^{\text{text}}(.)$	*14.54*	*6.73*
	$\text{M}(\mathcal{C}_{S_H \cup T_H}, \mathcal{V})$	$P(\mathcal{C}_{S_H \cup T_H}\|.)P(\mathcal{V}\|.)P(.)$	*7.57*	*2.53*
	True Prior	$P(\mathcal{V}\|.)P^{\star}(.)$	*4.92*	*2.63*
	Visual Bayes	$P^{\text{im}}(\mathcal{C}\|.)P(\mathcal{V}\|.)P^{\star}(.)$	*3.40*	*2.11*
Baselines	$\text{DeViSE}(\mathcal{V})$	$P(.\|\mathcal{V})$	10.73	3.62
	$\text{DeViSE}(\mathcal{C}_{S_H} \oplus \mathcal{V})$	$P(.\|\mathcal{C}_{S_H}, \mathcal{V})$	10.11	3.11
	$\text{M}(\mathcal{C}_{S_H} \oplus \mathcal{V})$	$P(\mathcal{C}_{S_H}, \mathcal{V}\|.)P(.)$	10.07	1.85
	$\text{M}(\mathcal{C}_I, \mathcal{V})$	$P(\mathcal{C}_I\|.)P(\mathcal{V}\|.)P(.)$	9.19	2.13
Our models	$\text{M}(\mathcal{V})$	$P(\mathcal{V}\|.)P(.)$	10.72	2.64
	$\text{M}(\mathcal{C}_{S_L}, \mathcal{V})$	$P(\mathcal{C}_{S_L}\|.)P(\mathcal{V}\|.)P(.)$	9.01	2.05
	$\text{M}(\mathcal{C}_{T_L}, \mathcal{V})$	$P(\mathcal{C}_{T_L}\|.)P(\mathcal{V}\|.)P(.)$	9.00	2.13
	$\text{M}(\mathcal{C}_{S_H}, \mathcal{V})$	$P(\mathcal{C}_{S_H}\|.)P(\mathcal{V}\|.)P(.)$	8.96	1.92
	$\text{M}(\mathcal{C}_{S_L \cup T_L}, \mathcal{V})$	$P(\mathcal{C}_{S_L \cup T_L}\|.)P(\mathcal{V}\|.)P(.)$	8.60	1.93
	$\text{M}(\mathcal{C}_{S_H \cup T_L}, \mathcal{V})$	$P(\mathcal{C}_{S_H \cup T_L}\|.)P(\mathcal{V}\|.)P(.)$	8.52	1.86
	$\text{M}(\mathcal{C}_{S_H \cup S_L \cup T_L}, \mathcal{V})$	$P(\mathcal{C}_{S_H \cup S_L \cup T_L}\|.)P(\mathcal{V}\|.)P(.)$	**8.31**	**1.79**

Table 3.4 – **Evaluation of baselines, scenarios and oracles** MFR performances (given in %). $p_{\text{sup}} = 50\%$. Oracle results, written in italics, are not taken into account to determine the best scores, written in bold.

3.5.2 Modeling contextual information

In this section, we compare the different context models; results are reported in Table 3.4.

First, underlying hypotheses of our model are experimentally tested:

- Modeling context and prior information with semantic representations (models $\text{M}(\mathcal{C}_\star, \mathcal{V})$) is far more efficient than using direct textual co-occurrences, as shown by the *Textual Bayes* baseline, which is the weaker model despite being an Oracle.

- We show that the hypothesis on the conditional independence of $\mathcal{C}$ and $\mathcal{V}$ is acceptable, as separately modeling $\mathcal{C}$ and $\mathcal{V}$ gives better results than jointly modeling them (i.e. $\text{M}(\mathcal{C}_{S_H}, \mathcal{V}) \overset{\text{MFR}}{<} \text{M}(\mathcal{C}_{S_H} \oplus \mathcal{V})$).

- We observe that our approach $\text{M}(\mathcal{V})$ is more efficient to capture the imbalanced class distribution of the source domain, compared to $\text{DeViSE}(\mathcal{V})$; indeed, on $\mathcal{S}$, *True Prior* $\approx \text{M}(\mathcal{V})$ (2.63 vs 2.64), whereas *True Prior* $\overset{\text{MFR}}{<} \text{DeViSE}(\mathcal{V})$ (2.63 vs 3.62). Even if the improvement is only significant for the source domain $\mathcal{S}$, it indicates that separately using information sources is clearly a superior approach to further integrate contextual information.

Second, as observed in the case of the context model S_H (Section 3.5.1), using contextual information is always beneficial. Indeed, all models with context $\mathrm{M}(\mathcal{C}_\star, \mathcal{V})$ improve over $\mathrm{M}(\mathcal{V})$ — which is the model with no contextual information — both on target and source domains. In more details, we observe that performances increase when additional information is used:

- when the bounding boxes annotations are available: all of our models that use both $\mathcal{C}$ and $\mathcal{V}$ outperform the baseline $\mathrm{M}(\mathcal{C}_I, \mathcal{V})$, which could also be explained by the useless noise outside the object boxes in the image and the difficulty of computing a global context from raw image,
- when context objects are labeled and high-level features are used instead of low-level features, e.g. $S_H \overset{\text{MFR}}{<} S_L$ and $S_H \cup T_H \overset{\text{MFR}}{<} S_H \cup T_L$,
- when more context objects are considered (e.g. $S_L \cup T_L \overset{\text{MFR}}{<} S_L$),
- when low-level information is used complementarily to high-level information (e.g. $S_L \cup S_H \cup T_L \overset{\text{MFR}}{<} S_L \cup T_L$). As a result, the best performance is attained for $\mathrm{M}(\mathcal{C}_{S_L \cup S_H \cup T_L}, \mathcal{V})$, with a 22% (resp. 32%) relative improvement in the target (resp. source) domain compared to $\mathrm{M}(\mathcal{V})$.

We note that there is still room for improvement to approach ground-truth distributions for objects of the target domain (e.g, towards word embeddings able to better capture visual context). Indeed, even if our models outperform *True Prior* and *Visual Bayes* on the source domain, these Oracle baselines are still better on the target domain, hence showing that learning the visual context of objects from textual data is challenging.

3.5.3 Qualitative Experiments

To gain a deeper understanding of contextual information, we compare in Figure 3.3 the predictions of $\mathrm{M}(\mathcal{V})$ and the global model $\mathrm{M}(\mathcal{C}, \mathcal{V})$. We randomly select five classes of the target domain and plot, for all instances of these classes in the test set of Visual Genome, the distribution of the predicted ranks of the correct class (in percentage); we also list the classes that appear the most in the context of these classes. We observe that, for certain classes (*player*, *handle* and *field*), contextual information helps to refine the predictions; for others (*house* and *dirt*), contextual information degrades the quality of the predictions.

First, we can outline that visual context can guide the model towards a more precise prediction. For example, a *player*, without context, could be categorized as *person*, *man* or *woman*; but visual context provides important complementary information (e.g, *helmet*, *baseball*) that grounds *person* in a sport setting, and thus suggests that the *person* could be playing. Visual context is also particularly

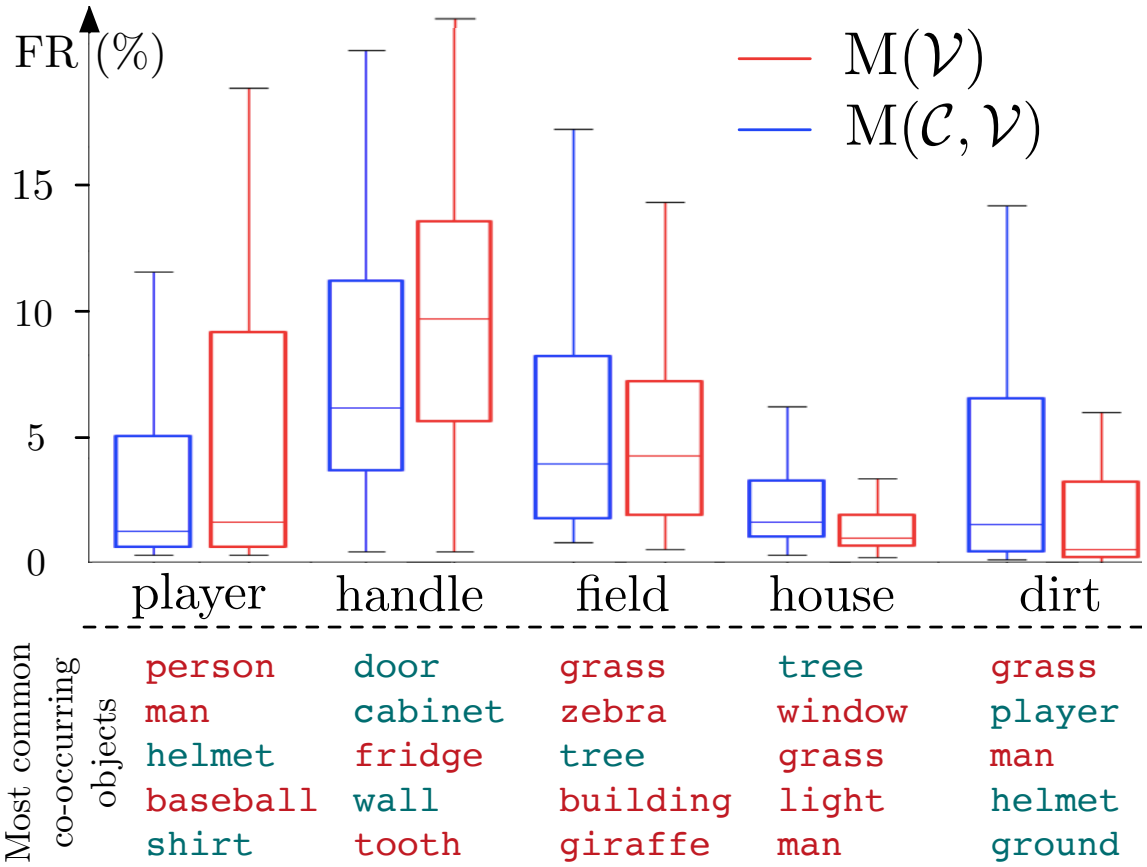

Figure 3.3 – Boxplot representing the distribution of the correct ranks (First Relevant in %) for five randomly selected classes of the target domain, with the context model $S_L \cup S_H \cup T_L$. Below are listed, by order of frequency, the classes that co-occur the most with the object of interest (classes of $\mathcal{T}$ in green; $\mathcal{S}$ in red).

relevant when the object of interest has a generic shape. For example, *handle*, without context, is visually similar to many round objects; but the presence of objects like *door* or *fridge* in the context helps determine the nature of the object of interest.

To get a better insight on the role of context, we cherry-picked examples where the visual or the prior component is inacurrate and the context component is able to counterbalance the final prediction (Figure 3.4). In (i), for example, the visual component ranks *flower* at position 223. However, the context component assesses *flower* to be highly probable in this context, due to the presence of source objects like *vase*, *water*, *stems* or *grass*, but also target objects like the other flowers around. At the inference phase, probabilities are aggregated and *flower* is ranked first.

It is worth noting that our work is not without limitations. Indeed, some classes (such as *house* and *dirt*) have a wide range of possible contexts; in these cases,

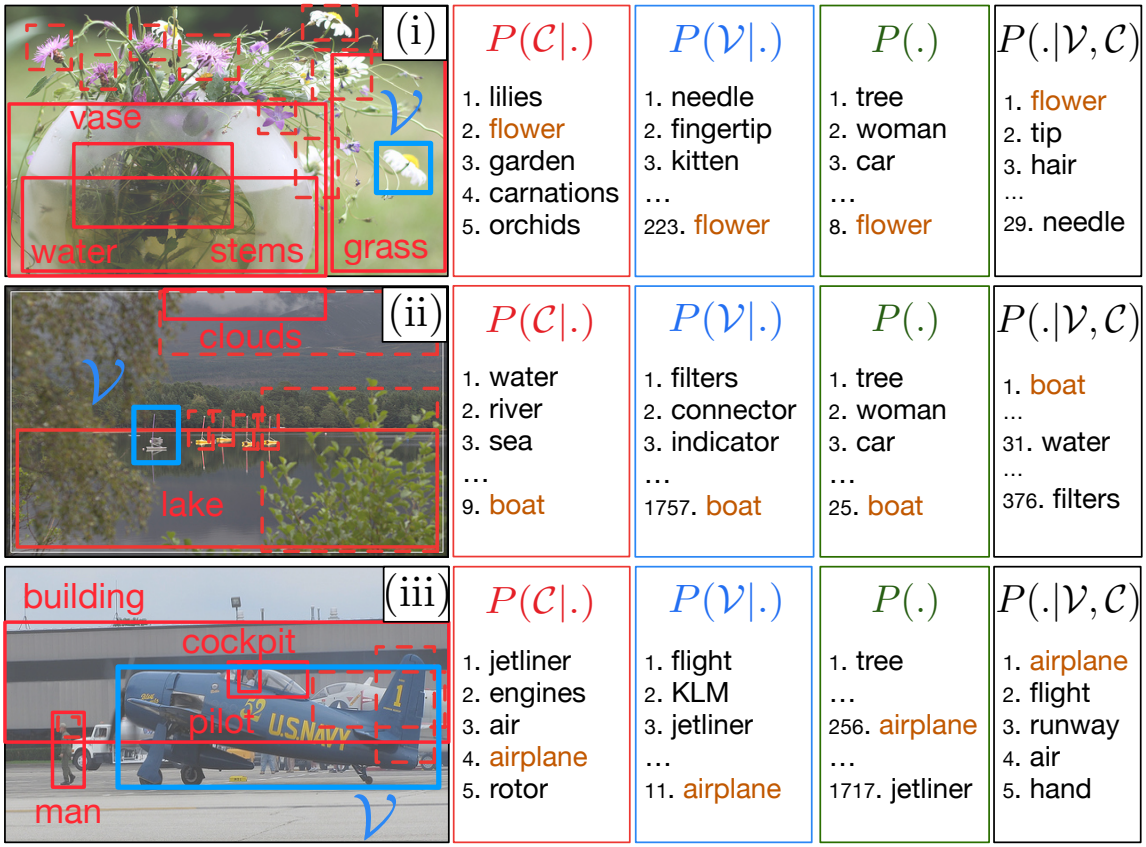

Figure 3.4 – Qualitative examples where the global model $\mathrm{M}(\mathcal{C}_{S_L \cup S_H \cup T_L}, \mathcal{V})$ correctly retrieves the class ($\mathcal{T}$ classes only).

context is not a discriminating factor. This is confirmed by a complementary analysis: the Spearman correlation between the number of unique context objects and δ_C, the relative gain of $\mathrm{M}(\mathcal{C}_{S_H}, \mathcal{V})$ over $\mathrm{M}(\mathcal{V})$, is $\rho = -0.31$. In other terms, contextual information is useful for specific objects, which appear in particular contexts; for objects that are too generic, adding contextual information can be a source of noise. This suggests, as an extension of the model, to add a specific module indicating whether contextual information should be used (based, for example, on the prediction of the prior module about the frequence of the object).

Additional examples are given in Figure 3.5, when an object occurs in an environment in which it is unexpected. For example, we have a picture of a kitchen where the object of interest to be predicted is "books". Given only the surrounding environment, predicted objects are logically related to the environment of a kitchen ("freezer", "oven", ...), and the correct label is badly ranked (because it is unexpected in such an environment). However, the model $\mathrm{M}(\mathcal{V})$ retrieves the correct label, given only the region of interest. Finally, integrating contextual information in the final model $\mathrm{M}(\mathcal{C}_{S_L \cup S_H \cup T_L}, \mathcal{V})$ leads to worse performances over $\mathrm{M}(\mathcal{V})$.

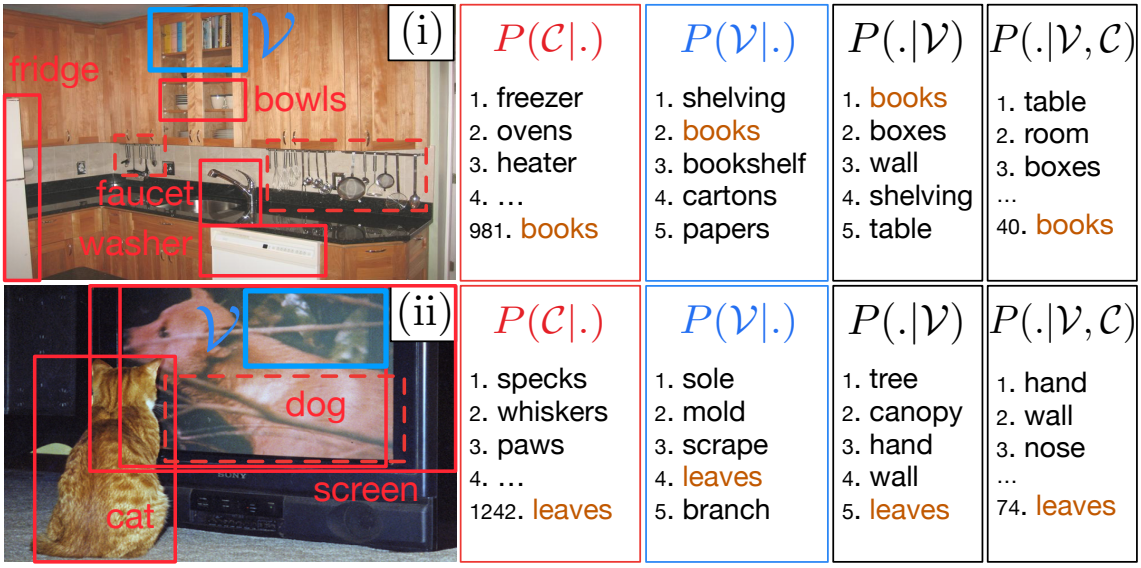

Figure 3.5 – Qualitative analysis: negative examples where the use of the context leads to degraded predictions, i.e. examples where model $\mathrm{M}(\mathcal{C}_{S_L \cup S_H \cup T_L}, \mathcal{V})$ is worse than the simpler model $\mathrm{M}(\mathcal{V})$ ($\mathcal{T}$ classes only).

3.6 Conclusion and Perspectives

3.6.1 Summary of the contributions

In this paper, we introduced a new approach for ZSL: *context-aware ZSL*, using data from Visual Genome. The goal is to determine the class of an object — delimited with a bounding box in an image — (that may not be known to the model) by using the complementary information of the visual context around this object. We proposed a corresponding model, based on three components: (i) a *visual* component that processes the visual appearance of objects, (ii) a *contextual* component that leverages the other objects of the image and (iii) a *prior* component that considers the frequence of objects.

We demonstrated experimentally that using this complementary contextual information enables to improve the performances of a ZSL model — in our case, the standard DeViSE model. We also showed that word representations contain information about the visual context of objects, and, more surprisingly, about the frequence of objects in images. Since our model is modular, we can interpret the contributions of each of the three components, and we provide qualitative analyzes to show examples where contextual information is useful, or not.

We thus provide the following answers to the Chapter Questions:

- **CQ1**: Textual representations, like Word2vec, encode information about the visual appearance, the visual context and visual frequence of objects.
- **CQ2**: Cross-modal models gain from being separated in modules, each focused on a distinct visual aspect. Bayesian modeling is an example of how to combine predictions made by model's components.

3.6.2 Perspectives

In this section, we present research perspectives following the contribution made in this Chapter.

Context-aware word embeddings In this Chapter, we use Word2Vec (Mikolov et al. 2013b) vectors for word representations. We show that these vectors encode some information about (i) the visual appearance of objects and (ii) their visual context in images.

A first extension would be to separately learn *grounded* word embeddings and replace Word2vec vectors by these grounded vectors. With such word vectors, the model may show higher performances for the visual component of our model. To learn grounded word embeddings, one must select a group of visual words, for which images are available during training. In our case, we would consider seen class labels, as unseen class labels are supposed to be unknown to the model. To learn such grounded representations, we could use for example, the *imagined* sequential model of Collell et al. 2017 or the Multi-Modal Skip-Gram of A. Lazaridou et al. 2015a. More interestingly, we could use the model presented in Eloi Zablocki et al. 2018a to learn word representations grounded *in context*. Indeed, we could learn such grounded vectors using the contextual information around objects from seen classes. With such word vectors, the model may show higher performances for the contextual component.

Another extension, would be to use our context-aware zero-shot learning model to learn grounded word representations. Such representations may carry information about the visual appearance of objects and contextual information. To do so, we would optimize jointly a Word2Vec loss and the three losses (visual, contextual and prior components) of our model.

Build a complete Zero-Shot Object Detection framework In the present Chapter, we suppose that the bounding boxes around objects are given, and our model uses contextual information around object to determine the class of an unknown object in the image. Considering a setting where bounding boxes are not given would be interesting, and would correspond to a Zero-Shot Object Detection task, as the goal would be to (i) find the objects in the image, and (ii) determining their class, using their visual appearance and contextual information.

(i) could be tackled using a Region Proposal Network (RPN), as done in the Faster R-CNN model (S. Ren et al. 2017). To do (ii), we would need to use our model in a Generalized ZSL setting, as objects can be either seen or unseen. Moreover, since there is no ground-truth annotations about the classes of objects, predictions should be made using the visual appearance of objects and low-level contextual information, i.e. model $\mathrm{M}(\mathcal{C}_{\mathcal{S}_L} \cup \mathcal{C}_{\mathcal{S}_T}, \mathcal{V})$.

CHAPTER 4

LEVERAGING WEAK/NON-EXISTENT CROSS-MODAL SUPERVISION

Contents

Chapter abstract

In Computer Vision, Zero-Shot Learning (ZSL) aims at classifying unseen *classes — classes for which no matching training image exists. Most of ZSL works learn a cross-modal mapping between images and class labels for seen classes. However, the data distribution of seen and unseen classes might differ, causing a domain shift problem. Following this observation, transductive ZSL (T-ZSL) assumes that unseen classes and their associated images are known*

during training, but not their correspondence. As current T-ZSL approaches do not scale efficiently when the number of seen classes is high, we tackle this problem with a new model for T-ZSL based upon CycleGAN. Our model jointly (i) projects images on their seen class labels with a supervised objective and (ii) aligns unseen class labels and visual exemplars with adversarial and cycle-consistency objectives. We show the efficiency of CM-GAN on the ImageNet T-ZSL task where we obtain state-of-the-art results. We further validate CM-GAN on a language grounding task, and on a new task that we propose: zero-shot sentence-to-image matching on MS COCO.

This work is currently under review at the Pattern Recognition Journal:

- *Patrick Bordes, Eloi Zablocki, Benjamin Piwowarski, Patrick Gallinari: "Transductive Zero-Shot Learning using Cross-Modal CycleGAN"*

4.1 Introduction

4.1.1 Positioning

In this Chapter, we present a contribution on the exploitation of weak/non-existent cross-modal alignment in multimodal tasks, either *NLP tasks aided by CV*, *CV tasks aided by NLP* or *Cross-Modal tasks*. Thus, this Chapter is linked to the three Research Questions defined in the Introduction of the present thesis.

In the vast majority of cases (see Background Chapter), a *direct supervision* between modalities is exploited in multimodal tasks. In this Chapter, we adopt a new point of view: exploiting *unsupervised* relationships between modalities. We focus on the Zero-Shot Learning task, but we also consider the Cross-Modal retrieval task and the Visual Grounding of Language task in the experimental section (Section 4.4), thus covering the three groups of multimodal tasks defined in the Introduction. We thus formulate a Chapter Question (CQ), that we strive to answer throughout this Chapter:

- **CQ**: Can multimodal tasks benefit from multimodal data when the cross-modal supervision is weak, or non-existent ?

4.1.2 Transductive Zero-Shot Learning

As pointed out in Fu et al. 2015a, standard Zero-Shot Learning (ZSL) models (Section 2.3.1) lead to a *domain shift* problem: since the sets of seen and unseen classes are disjoint and potentially unrelated, the learned projection functions are biased towards seen classes. Transductive Zero-Shot Learning (T-ZSL) (D. Zhou et al. 2003; Wan et al. 2019) attemps to solve this domain shift problem. In T-ZSL, the unlabeled images corresponding to unseen classes are available during

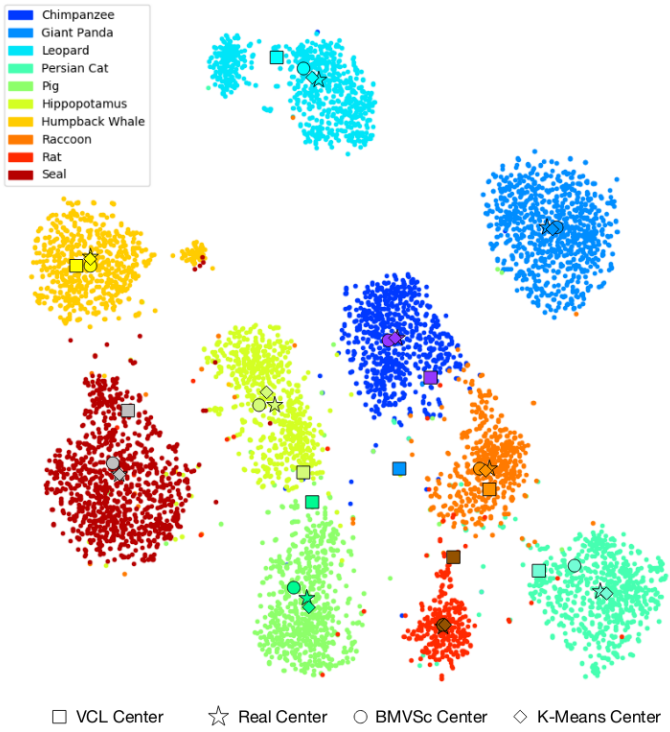

Figure 4.1 – Transductive Zero-Shot Learning with Visual Structure Constraint. Illustration taken from Wan et al. 2019.

training (Verma et al. 2017; Ye et al. 2017; Wan et al. 2019). A wide variety of T-ZSL approaches have been proposed.

Some methods use label propagation (Fujiwara et al. 2014); for example, (Ye et al. 2017) "rectify" a prediction matrix obtained by traditional ZSL (matrix of cosine similarities between all images and all classes) to cope with the domain-shift using an affinity matrix that measures semantic distances between all classes. Other methods, such as DIPL (A. Zhao et al. 2018), project visual features corresponding to unseen classes close to the closest unseen class representation with a min-min optimization problem. Another set of methods exploit the natural clusters in the visual embedding space, such as (Wan et al. 2019) who use a K-means clustering algorithm to determine the centroids of images corresponding to unseen classes. This approach is illustrated in Figure 4.1, where Convolutional Neural Network (CNN) features of images corresponding to the 10 unseen classes of the AwA2 dataset (Yongqin Xian et al. 2019) are visualized using t-SNE (Maaten et al. 2008); stars correspond to real clusters, other shapes are centers predicted by various versions of the proposed model. Figure 4.1 illustrates that the visual space is well-clustered for simple datasets like AwA2 that have a low number of unseen classes.

However, current models assume the visual space to be well-clustered, which is not the case when the number of classes is high, thus explaining why these

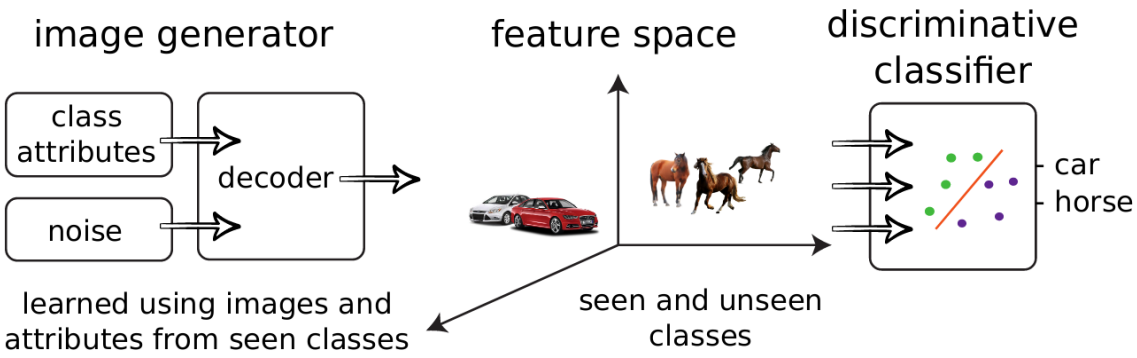

Figure 4.2 – Generating Visual Representations for Zero-Shot Classification. Illustration taken from Jurie et al. 2017.

methods fail to benefit from the transductive setting. Transductive ZSL is challenging on ImageNet (20K unseen classes) due to the high number of classes. Indeed, the 20K*20K affinity matrix in Ye et al. 2017 is impossible to invert in label propagation; well-separatedness of unseen classes in the visual embedding space is not verified in (Wan et al. 2019), and min-min optimization is extremely long and only brings additional noise in (A. Zhao et al. 2018), as we show in the experiments. In the present thesis, we tackle this challenge using a Cross-Modal CycleGAN model: this is the purpose of Chapter 4.

4.1.2.1 Generative models in ZSL

Another body of work expresses visual features as probability distributions written conditionally to label representations (Wenlin Wang et al. 2018; Jurie et al. 2017; Mishra et al. 2018) using generative models such as Variational Auto-Encoder (VAE) (Kingma et al. 2014b) or Generative Adversarial Network (GAN). As for previously described approaches, they first rely on a semantic representation of labels (e.g. Word2Vec representation of the class). A label representation is used to define a conditional probability distribution over the space of images, or more precisely of image representations.

Some works generate visual exemplars for unseen classes, that are then fed to a supervised model, as in Gvf (Jurie et al. 2017) with GANs — illustrated in Figure 4.2 — or (Mishra et al. 2018) with VAEs. In Xian et al. 2018, the generation of features and the training of the supervised model is done simultaneously, in an end-to-end fashion.

Other approaches predict the correct label using a Bayesian approach (Khare et al. 2019), or by maximizing the variational lower bound of a VAE (Wenlin Wang et al. 2018). Generative models are usually harder to learn (since they need to model the visual distribution) and perform worse than their discriminative counterparts – but they can be used in a transductive ZSL setting.

4.1.3 Contributions

In the present paper, we tackle transductive ZSL with a high number of unseen classes, by building upon CycleGan (J. Zhu et al. 2017): image and label distributions are aligned with an adversarial loss while ensuring that the structure of both spaces is preserved with a cycle-consistency loss. Thus, useful information is learned from the unseen classes distribution. Our work follows unsupervised translation works that proved efficient on large-scale uni-modal datasets (G. Lample et al. 2018), and extends them to a cross-modal setting where

- (P1) the geometry of the visual and textual spaces are essentially different
- (P2) there is a high imbalance between the number of images and the number of class labels.

In the model that we propose, called Cross-Modal CycleGAN (CM-GAN), two cross-modal mappings (text-to-image and image-to-text) are learned between a visual space and a textual space using:

- (i) a *supervised* max-margin triplet objective trained on seen classes,
- (ii) an *unsupervised* CycleGAN objective trained on unseen classes.

We tackle (P1) by representing images as linear combinations of class labels representations, following the CONSE (Norouzi et al. 2014) approach. This enables textual and visual distributions to be close and thus adversarial losses to be learned. We tackle (P2) by using an alternating optimization scheme where we progressively refine the projection of images and labels.

We show that CM-GAN is successful when the number of unseen classes is high (Section 4.4.1), namely on the ImageNet T-ZSL task where it achieves SOTA results. We further validate the efficiency of CM-GAN on a language grounding task (Section 4.4.2) and on a new task, namely *zero-shot image-to-sentence matching*, on MS COCO (Section 4.4.3).

4.2 The Cross-Modal CycleGAN Model

Adversarial learning aims at estimating a mapping between two data distributions, from non-aligned data. In word-to-word translation, (G. Lample et al. 2018) learn to align two unpaired word spaces from different languages using an adversarial loss. In image-to-image translation, CycleGAN (J. Zhu et al. 2017) has been widely adopted (Y. Lu et al. 2017; Almahairi et al. 2019; J. Zhao et al. 2019); it adds to the adversarial objectives a cycle-consistency objective to constrain mappings to be somewhat invertible. Despite being widely used with uni-modal data, CycleGAN has rarely been applied to cross-modal translation, with some

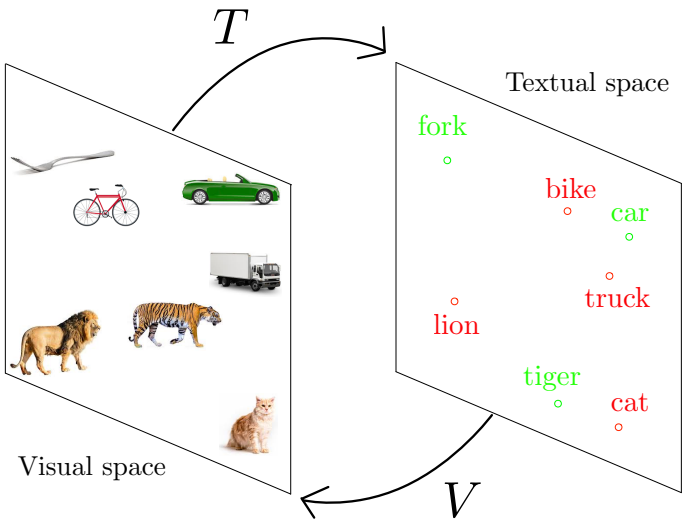

Figure 4.3 – Model overview; our goal is to learn two mapping functions T and V between a visual space $\mathcal{V}$ and a textual space $\mathcal{T}$, where class labels are either *seen* (in red) or *unseen* (in green).

exceptions such as speech-to-text alignment (Chung et al. 2018). However, unlike speech and text, image and text have different semantics. To cope with this problem, our model represents images with CONSE embeddings.

In this section, we present our CM-GAN model for Transductive Zero-Shot Learning, illustrated in Figure 4.3. Images are embedded in a visual space $\mathcal{V}$, and class labels are embedded in a textual space $\mathcal{T}$. Our goal is to learn cross-modal functions T and V such that T projects images to their corresponding class labels and V projects class labels to their corresponding images. Once such functions have been learned, retrieval is made indifferently in $\mathcal{V}$ or $\mathcal{T}$ using a nearest neighbor search. Due to the ZSL setting, labels either belong to the seen classes $\mathcal{S}_\mathcal{T}$ (marked in red in Figure 4.3) or unseen classes $\mathcal{U}_\mathcal{T}$ (green). We note $\mathcal{S}_\mathcal{V}$ and $\mathcal{U}_\mathcal{V}$ their corresponding images.

4.2.1 Data Representations

The first issue that needs to be addressed is to embed modalities so that textual and visual distributions are somewhat close and admit a meaningful mapping, as in (G. Lample et al. 2018) where two unpaired Word2vec spaces from distinct languages are aligned using adversarial losses. We present below our choices for text and image representations.

Class labels representation Classes are represented with the Skip-Gram embedding of their label (Mikolov et al. 2013b). We call T_0 the Word2vec function, trained on Wikipedia, that takes a word as input and outputs a vector of $\mathbb{R}^d$,

with $d = 500$. If the textual space contains sentences S instead of words, they are represented by the sum of their word embeddings:

$$s = \sum_{w \in S} T_0(w) \tag{4.1}$$

Image representation We call V_0 the image embedding function, which transforms an input image into a vector of $\mathbb{R}^d$. We argue that visual representations should be "homogeneous" to the classes embeddings so that distributions can be mapped on one another. Thus, we propose to use the CONSE model (Norouzi et al. 2014), in which an image is represented as a convex combination of class label embeddings. More precisely, for a given image, let us call $p(i)$ the distribution over the seen classes $i \in [1, |\mathcal{S}_\mathcal{T}|]$ output by the CNN, and T_K the indices of the K most probable classes. The representation of an image is then:

$$V_0(v) = \sum_{i \in T_K} p(i|i \in T_K) T_0(w_i) \tag{4.2}$$

where w_i is the label of class i and $p(i|i \in T_K) \propto p(i)$.

To assess our intuition that CONSE leads to a visual space that is semantically close to the textual (Word2vec) modality, we conduct a preliminary experiment by comparing CONSE to other visual representations, namely CNN (K. He et al. 2016) and DeViSE (Frome et al. 2013), and report results in Table 4.1. To do so, we use the ρ_{vis} metric, that we define in Chapter 6, which measures the similarity of two sets of vectors even if they do not share a joint embedding space. More precisely, we set:

$$\rho_{vis}(\mathcal{S}) = \rho(cos(t, t'), cos(v, v')) \tag{4.3}$$

where ρ is the Pearson correlation and pairs v, t and v', t' are aligned and sampled from $\mathcal{S}$; similarly, $\rho_{vis}(\mathcal{U})$ can be defined when pairs are sampled from unseen data. The discrepancy between $\mathcal{S}$ and $\mathcal{U}$ is due to the fact that the seen set $\mathcal{S}$ is used as supervision to learn the various models of Table 4.1; thus, modalities are better aligned for seen classes. We observe that CONSE leads to a better similarity between modalities, which confirms preliminary results where adversarial losses failed to produce meaningful models when applied to CNN or DeViSE vectors.

Having set the initial representations of classes and images, we now proceed to define the learning objectives.

4.2.2 Supervised Loss

The *supervised loss* leverages the information of seen classes, as illustrated in Figure 4.4. The correspondence between images and seen class labels is a many-

Model	$\mathcal{S}$	$\mathcal{U}$
Random	11.6	11.6
CNN K. He et al. 2016	30.7	29.4
DeViSE Frome et al. 2013	53.8	36.9
CONSE Norouzi et al. 2014	**92.1**	**45.9**

Table 4.1 – Preliminary experiment: comparison of visual representations. The metric is ρ_{vis} (in %) computed on ImageNet's $\mathcal{S}$ (1K seen classes) and $\mathcal{U}$ (20K unseen classes) — the higher the better. CONSE and DeViSE were re-implemented.

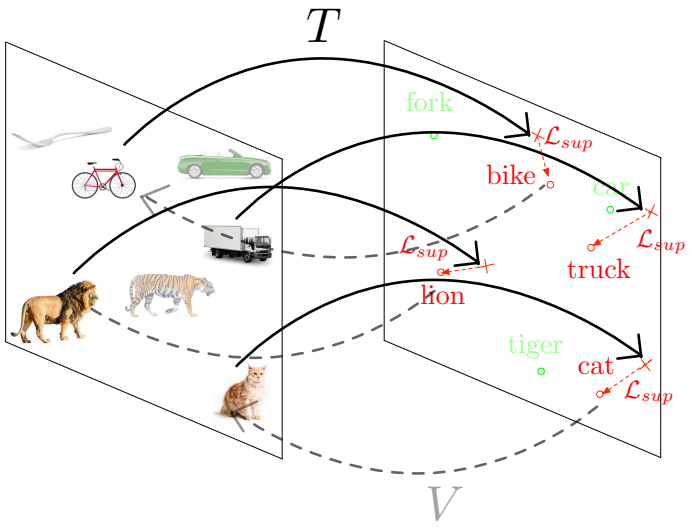

Figure 4.4 – The *supervised objective* learns to projects images (resp. class labels) to their corresponding class labels (resp. images) for seen classes.

to-one correspondence, that can be exploited using a standard max-margin triplet loss $\mathcal{L}_{sup}$ — commonly used (Frome et al. 2013) to bring closer elements from distinct modalities in a common space:

$$\mathcal{L}_{sup} = \mathop{\mathbb{E}}_{v,t,v^-,t^-} \big(\lfloor \gamma - \cos(T(\hat{v}), \hat{t}) + \cos(T(\hat{v}), \hat{t}^-) \rfloor_+ + \lfloor \gamma - \cos(\hat{v}, V(\hat{t})) + \cos(\hat{v}^-, V(\hat{t})) \rfloor_+ \big) \quad (4.4)$$

where v is an image with t its label sampled from the seen images, and v^- and t^- are negative examples sampled from the set of images and labels respectively. We denote $\hat{v}$ (resp. $\hat{t}$) the current representation of the image v (resp. of label t) that we define precisely in Section 4.2.4. The margin γ is set to 0.5.

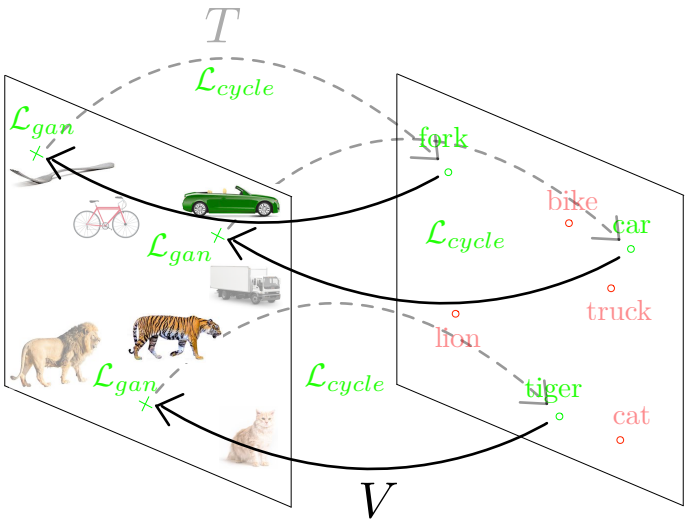

Figure 4.5 – The *unsupervised objective* is a CycleGAN objective used in the unseen classes to align the visual and textual distributions.

4.2.3 Adversarial and Cycle-Consistency Losses

The *transductive/unsupervised loss* aims at capturing information from the unseen classes, as illustrated in Figure 4.5. Since there is no known mapping between unseen class labels and their corresponding images, we can only use the distribution of images and classes to align modalities. We use CycleGan (J. Zhu et al. 2017), which has proven useful in the unpaired image-to-image translation task. The learned mappings are the cross-modal functions T and V, and discriminators in each modality $D_{\mathcal{V}}$ and $D_{\mathcal{T}}$. The CycleGan objective consists of (i) adversarial losses in both spaces $\mathcal{L}^v_{gan}$ and $\mathcal{L}^t_{gan}$ to align the textual and visual distributions and (ii) a cycle-consistency loss $\mathcal{L}_c$ weighted by a scalar λ_c to ensure that the mappings are somewhat reversible:

$$\mathcal{L}_{cgan} = \mathcal{L}^v_{gan} + \mathcal{L}^t_{gan} + \lambda_c.\mathcal{L}_c \tag{4.5}$$

The individual losses write as follows:

$$\mathcal{L}^v_{gan} = \mathop{\mathbb{E}}_{t\in\mathcal{U}_{\mathcal{T}}}[\log D_{\mathcal{T}}(\hat{t})] + \mathop{\mathbb{E}}_{v\in\mathcal{U}_{\mathcal{V}}}[\log(1 - D_{\mathcal{T}}(T(\hat{v})))] \tag{4.6}$$

$$\mathcal{L}^t_{gan} = \mathop{\mathbb{E}}_{v\in\mathcal{U}_{\mathcal{V}}}[\log D_{\mathcal{V}}(\hat{v})] + \mathop{\mathbb{E}}_{t\in\mathcal{U}_{\mathcal{T}}}[\log(1 - D_{\mathcal{V}}(V(\hat{t})))] \tag{4.7}$$

$$\mathcal{L}_c = \mathop{\mathbb{E}}_{t\in\mathcal{U}_{\mathcal{T}}}[\|T(V(\hat{t})) - \hat{t}\|_2] + \mathop{\mathbb{E}}_{v\in\mathcal{U}_{\mathcal{V}}}[\|V(T(\hat{v})) - \hat{v}\|_2] \tag{4.8}$$

where $\mathcal{U}_{\mathcal{T}}$ and $\mathcal{U}_{\mathcal{V}}$ are the textual and visual distributions for unseen classes. V aims at generating visual representations that are indistinguishable from vectors

of $\mathcal{U}_\mathcal{V}$, whereas the discriminator $D_\mathcal{V}$ aims at distinguishing elements from $\mathcal{U}_\mathcal{V}$ and elements from $V(\mathcal{U}_\mathcal{T})$. As in (Goodfellow et al. 2014), V aims at minimizing $\mathcal{L}^t_{GAN}$, and the discriminator $D_\mathcal{V}$ aims at maximizing it in an adversarial fashion.

4.2.4 Learning

To learn the cross-modal functions T and V, we adopt an alternate iterative learning procedure, as preliminary experiments showed that jointly optimizing $\mathcal{L}_{cgan}$ and $\mathcal{L}_{sup}$ leads to highly unstable training. Instead of learning T and V in one step, we refine these mappings iteratively by composing the functions T_k and V_k learned at each optimization step k, giving rise to the global mappings $\hat{T}_k$ and $\hat{V}_k$. Thus, the supervised and unsupervised losses (Equation 4.4 and Equation 4.5) are optimized using data from the previous step: $\hat{v} = \hat{T}_{k-1}(v)$ for each image v and $\hat{t} = \hat{V}_{k-1}(t)$ for each class label t; these representations are fixed to avoid over-fitting.

Supervised step During a supervised step k, we optimize T_k and V_k – modeled as 2-layer Perceptrons – with Equation 4.4. We notice experimentally that our validation measure (Section 4.3.4) is optimal when retrieval is performed in the textual space $\mathcal{T}$, probably due to the many-to-one mapping that provides T_k with significantly more training data than V_k as input (1.3M images vs 1K seen classes). Thus, we do not use the learned V_k, and we set $\hat{T}_k = T_k \circ \hat{T}_{k-1}$ and $\hat{V}_k = \hat{V}_{k-1}$.

Unsupervised/transductive step During an unsupervised step k, we optimize T_k and V_k – modeled as 2-layer Perceptrons – with Equation 4.5. We notice experimentally that our validation measure is optimal when retrieval is performed in the visual space $\mathcal{V}$, probably because the discriminator $D_\mathcal{V}$ has significantly more training data on which to be trained (13M images vs 20K unseen classes), thus leading to a better cross-modal V_k. Consequently, we don't use the learned T_k, and we set $\hat{T}_k = \hat{T}_{k-1}$ and $\hat{V}_k = V_k \circ \hat{V}_{k-1}$.

Outcome When the unsupervised validation criterion (see Section 4.3.4) shows no improvement at the end of a step K, training is stopped, and we obtain the final functions $\hat{V}_K$ and $\hat{T}_K$.

4.3 Experimental Protocol

4.3.1 Datasets

ImageNet ImageNet (Farhadi et al. 2009a) consists of 14.2M images, corresponding to 21841 classes, 1000 of which are seen classes, and the rest unseen. Among all unseen classes, we keep 20345 classes that have a Word2vec embedding (compound words are averaged); we used the same word embeddings and same classes as (Changpinyo et al. 2016). We note *ImageNet-Full* the dataset with all 20345 unseen classes. Among these classes, *2-hop* (resp. *3-hop*) consists of 1509 (resp. 7678) classes within two (resp. three) hops of a seen class in WordNet — it is thought that classes further away from seen ones (e.g. *3-hop*) are harder to learn. We also consider *ImageNet-360*, which is widely adopted among the ZSL literature because it is substantially smaller and allows comparison with the literature (360 unseen classes, with 400K images).

MS COCO We use the MS COCO dataset (T. Lin et al. 2014b) to tackle a new task that we introduce in this paper, that we call *zero-shot sentence-to-image matching*. We suppose that no text-to-image correspondence is known, and we evaluate our model on cross-modal retrieval. This task is interesting as (i) no supervised information can be used, (ii) it features sentences instead of words, and (iii) it extends ZSL to a very high number of classes (as many classes as sentences). The training set consists of 118K images, with 5 captions per image. Evaluation is performed over 1K images (along with the corresponding 5K captions) from the test set of MS COCO.

4.3.2 Evaluation Metrics

In the experiments, we consider the two standard evaluation settings: **Zero-Shot Learning** (ZSL), in which the image label is searched among unseen classes $\mathcal{U}$; and the more challenging **Generalized Zero-Shot Learning** (G-ZSL) where the class is searched among seen and unseen classes.

Following (A. Zhao et al. 2018) , we use the Recall at rank $k \in \{1, 2, 5, 10, 20\}$ metric — noted R_k — defined as the percentage of images for which the correct label is present in the top k predictions of the model.

Furthermore, following (E. Zablocki et al. 2019), we use the Mean First Relevant (MFR) metric to evaluate our model scenarios, as this metric is more stable than R_k and is not sensitive to the number of classes, thus enabling fine-grained model comparisons. MFR is defined as the mean value of the rank of the correct class among the model's predictions, averaged over the set of test images $\mathcal{U}_\mathcal{V}$ and linearly re-scaled so that the random model has a 50% score:

$$\text{MFR} = \frac{100}{K.|\mathcal{U}_\mathcal{V}|} \sum_{v \in \mathcal{U}_\mathcal{V}} \text{FR}_v$$

where $K = |\mathcal{U}_\mathcal{T}|$ for ZSL and $K = |\mathcal{U}_\mathcal{T}| + |\mathcal{S}_\mathcal{T}|$ for G-ZSL.

4.3.3 Baselines

For ImageNet, we compare CM-GAN to the standard models **DeViSE** (Frome et al. 2013), **CONSE** (Norouzi et al. 2014), **SYNC** (Changpinyo et al. 2016) and **EXEM** (Changpinyo et al. 2017), to state-of-the-art models **Gvf** (Jurie et al. 2017) and **DIPL** (A. Zhao et al. 2018), and to VAE-based models **VZSL** (Wenlin Wang et al. 2018), **CVAE-ZSL** (Mishra et al. 2018) and **SE-ZSL** (Verma et al. 2018). All reported results are extracted from the original papers, which explains that we could not report all metrics for all models. For DIPL, SOTA on ImageNet-360, we re-implement their model to get ImageNet-Full scores.

For MS COCO, as our unsupervised setting is new in the literature, we only compare to a *supervised* baseline, for which we suppose that the sentence-to-image alignment is known and the $\mathcal{L}_{sup}$ objective is optimized to map CONSE representation to the corresponding sentence vectors.

In the experiments performed on MS COCO, our validation criterion is the unsupervised criterion described in (G. Lample et al. 2018): the mean cosine similarity between a set of images and their predicted sentences (from a selected 1K images/ 5K captions of the validation set). Due to the absence of supervised data, there is no iterative process, but only one unsupervised step where $\mathcal{L}_{cgan}$ is optimized.

4.3.4 Implementation Details

In the experiments performed on ImageNet-Full and ImageNet-360, our validation metric is the value of the max-margin triplet loss $\mathcal{L}_{sup}$ computed on the seen classes, to avoid the over-fitting of the unsupervised loss — we compute separately both rows of Equation 4.4 to determine in which space retrieval is optimal (cf Section 4.2.4). This metric is used to determine $\lambda_c \in \{1, 5, 10\}$ and the stopping step. Our final model is optimal at $K = 6$ steps. Selected parameters for the unsupervised steps are respectively $\lambda_c = 1, 10, 1$.

All images were processed with Inception-V3 (Szegedy et al. 2016a) to build CONSE embeddings.

4.4 Results

In this section, we perform extensive experiments to show the effectiveness of CM-GAN on the ImageNet T-ZSL task (Section 4.4.1) — on which a post-analysis on word embeddings is provided (Section 4.4.2) and on a task we introduce: *zero-shot sentence-to-image matching* on MS COCO (Section 4.4.3).

4.4.1 Zero-Shot Learning on ImageNet

Quantitative results on ImageNet-Full are provided in Table 4.2 for the ZSL and G-ZSL tasks, and results on ImageNet-360 are provided in Table 4.3 for ZSL.

For ZSL, CM-GAN generally outperforms all models on *All*, *3-hop* and *2-hop* — results on these benchmarks are increasing due to the rising visual and semantic proximity of the class labels to the seen class labels, confirming previous works findings. We notice that Gvf, which generates novel visual exemplars, shows better performances than models that learn linear or non-linear cross-modal projections (DeViSE, EXEM) or hybrid models such as CONSE or SYNC. On ImageNet-360, CM-GAN outperforms methods than rely on generative models and VAEs such as VZSL, CVAE-ZSL and SE-ZSL, thus proving that our approach based on CycleGAN captures interesting information from the distribution of unseen classes. We observe that CM-GAN outperforms DIPL* on ImageNet-Full, and DIPL outperforms CM-GAN on ImageNet-360. Indeed, DIPL's transductive loss aims at bringing closer visual features to the closest class label among unseen classes: while this method efficiently constrains the solution when the number of unseen classes is low (360), its nearest neighbor search over a large number of classes (20K) may bring additional noise that degrade ZSL results.

Comparing G-ZSL to ZSL allows to analyze whether a model has a tendency to predict seen classes first. For G-ZSL, because they are based on nearest neighbor search, CONSE, CONSE* and CM-GAN perform worse than Gvf (a classifier over seen and unseen classes) at low recall ranks. Interestingly, our re-implementation CONSE*, which is based on a better CNN than the original CONSE, is even more penalized. At higher ranks, we notice that CM-GAN eventually outperforms Gvf for G-ZSL, showing that the cross-modal functions are correctly learned, albeit with a slight overfit towards seen classes. This confirms findings made in (Yongqin Xian et al. 2019) for CONSE, which shares the same results tendencies than CM-GAN — expected as our model builds upon CONSE (Section 4.2.1).

To further analyze our model, we provide an ablation study in Table 4.4. We report models where losses $\mathcal{L}_c$, $\mathcal{L}_{gan} + \lambda_c.\mathcal{L}_c$, $\mathcal{L}_{gan}$ and $\mathcal{L}_{sup}$ are optimized individually. Init. corresponds to the initialization of our model, with CONSE embeddings as visual features; in other terms, using functions T_0 and V_0 defined in Section 4.2.1. We observe that:

		Model	R_1	R_2	R_5	R_{10}	R_{20}
Zero-Shot	All	CONSE	1.4	2.2	3.9	5.8	8.3
		CONSE*	1.76	2.8	4.77	7.07	10.04
		DeViSE	0.8	1.4	2.5	3.9	6.0
		SYNC	1.5	2.4	4.5	7.1	10.9
		DIPL*	1.47	2.52	4.82	7.59	11.52
		EXEM	1.8	2.9	5.3	8.2	12.2
		Gvf	1.90	3.03	5.67	8.31	**13.14**
		CM-GAN	**1.99**	**3.18**	**5.76**	**8.64**	12.57
	3_{hop}	CONSE	2.7	4.4	7.8	11.5	16.1
		CONSE*	3.43	5.22	8.96	12.96	18.21
		DeViSE	1.7	2.9	5.3	8.2	12.5
		SYNC	2.9	4.9	9.2	14.2	20.9
		DIPL*	2.81	5.02	9.87	15.64	22.1
		EXEM	3.6	5.9	10.7	16.1	23.1
		Gvf	3.58	5.97	11.03	16.51	**23.88**
		CM-GAN	**3.88**	**6.15**	**11.25**	**16.66**	23.4
	2_{hop}	CONSE	9.4	15.1	24.7	32.7	41.8
		CONSE*	10.67	15.58	25.24	34.29	45.31
		DeViSE	6.0	10.0	18.1	26.4	36.4
		SYNC	10.5	16.7	28.6	40.1	52.0
		DIPL*	10.46	16.79	28.23	39.4	52.08
		EXEM	12.5	19.5	32.3	43.7	55.2
		Gvf	13.05	**21.52**	33.71	43.91	**57.31**
		CM-GAN	**13.7**	20.96	**33.73**	**45.51**	56.31
Generalized Zero-Shot	All	CONSE	0.2	1.2	3.0	5.0	7.5
		CONSE*	0.13	1.41	3.62	5.97	8.93
		Gvf	**1.03**	**1.93**	**4.98**	6.23	10.26
		CM-GAN	0.16	1.5	4.24	**7.28**	**11.28**
	3_{hop}	CONSE	0.2	2.4	5.9	9.7	14.3
		CONSE*	0.21	2.65	6.76	10.77	16.01
		Gvf	**1.99**	**4.01**	6.74	11.72	16.34
		CM-GAN	0.26	2.65	**7.94**	**13.73**	**20.56**
	2_{hop}	CONSE	0.3	7.1	17.2	24.9	33.5
		CONSE*	0.17	6.3	16.37	24.67	34.45
		Gvf	**4.93**	**13.02**	20.81	31.48	45.31
		CM-GAN	0.18	7.06	**22.55**	**34.17**	**46.86**

Table 4.2 – ZSL and G-ZSL results on ImageNet-Full. Models marked with * were re-implemented.

Model	R_5
DeViSE Frome et al. 2013	12.8
ConSE Norouzi et al. 2014	15.5
VZSL Wenlin Wang et al. 2018	23.1
CVAE-ZSL Mishra et al. 2018	24.7
SE-ZSL Verma et al. 2018	25.4
DIPL A. Zhao et al. 2018	**31.7**
CM-GAN	25.9

Table 4.3 – ZSL results on ImageNet-360.

Model		2_{hop}	3_{hop}	All	2_{hop}	3_{hop}	All
Random		50	50	50	50	50	50
Init.	CONSE	8.88	12.37	15.35	9.88	12.71	15.65
cycle	$\mathcal{L}_c$	7.32	10.84	13.87	7.81	11.	14.3
gan	$\mathcal{L}_{gan}$	7.51	11.41	14.25	7.57	11.03	13.68
cgan	$\mathcal{L}_{gan} + \lambda_c.\mathcal{L}_c$	7.34	10.73	13.34	7.72	10.79	13.32
sup	$\mathcal{L}_{sup}$	5.83	8.88	11.07	6.06	8.87	11.13
CM-GAN		**5.17**	**8.23**	**9.98**	**5.3**	**8.17**	**9.97**

Table 4.4 – Ablation study. Measure: MFR (the lower the better).

- sup > cgan: the supervised step is better than the unsupervised step, which is expected since the alignment information is used;
- cgan > gan and cgan > cycle, which indicates that both cycle-consistency and adversarial losses are complementary;
- CM-GAN > sup and CM-GAN > cgan, which shows the benefits of the alternating learning process;
- our final model CM-GAN is the best scenario.

The complementarity of supervised and transductive objectives is illustrated in Figure 4.6: we randomly sample five unseen classes and visualize the evolution of their textual and visual representations at each model optimization step (Section 4.2.4): between even and uneven (resp. uneven and even) steps, visual centroids (resp. class labels) are moving. We observe that modalities align with each supervised (i,iii,v) and unsupervised (ii,iv,vi) iteration.

4.4.2 Learning Grounded Word Representations with CM-GAN

In this section, we provide a post-analysis of Section 4.4.1 and further validate CM-GAN by applying it to the *visual grounding of language* task (A. Lazaridou et al.

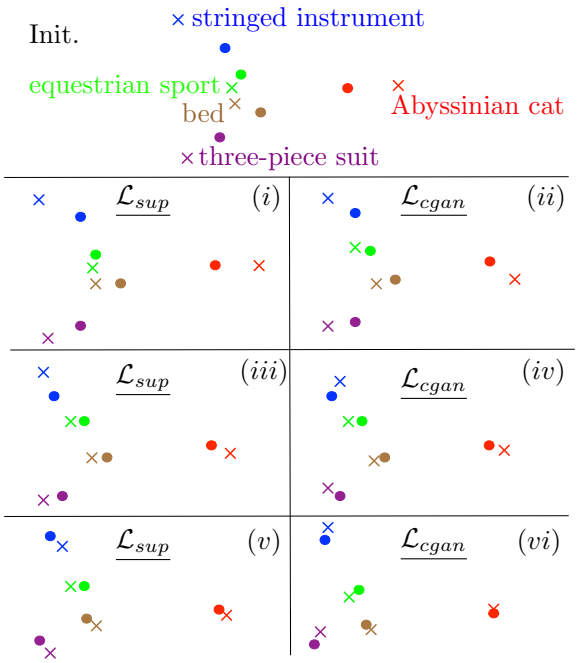

Figure 4.6 – PCA visualization of visual and textual data (represented in the same space) along with model's iterations for five randomly sampled unseen classes. Circle: centroid of visual features. Cross: class label.

Model	MEN	SemSim	SimLex	VisSim	WordSim	Conc.	Avg.
X	68	60	33	49	**62**	63	55.8
$V_{\text{sup}}(X)$	69	66	34	57	59	59	57.3
$X \oplus V_{\text{sup}}(X)$	70	**69**	36	**58**	59	58	58.3
$V_{\text{trans}}(X)$	70	62	35	51	**62**	62	57
$X \oplus V_{\text{trans}}(X)$	**72**	**69**	**37**	55	58	**64**	**59.2**
$V_{\text{sup}}(X) \oplus V_{\text{trans}}(X)$	70	**69**	36	**58**	59	58	58.3

Table 4.5 – Grounded vectors evaluation. Concatenated vectors were projected with PCA so that all vectors have the same dimension.

2015b), which aims at enhancing textual representations using visual information — see Section 2.2.1 of the Background Chapter for an overview of Visual G rounding of Language. Unlike previous models, that leverage direct correspondence between words and images, we also exploit unsupervised information. In this section, based on our CM-GAN model trained on ImageNet (Section 4.4.1), we learn grounded word representations using information from seen and unseen classes. For a word embedded as $X \in \mathbb{R}^{300}$ with Word2vec, we note $V_{\text{sup}}(X)$ (resp. $V_{\text{trans}}(X)$) the cross-modal projection of X learned using a supervised (resp. transductive) step. (Collell et al. 2017) has shown that X (*purely textual* as it is learned using Word2vec on textual corpus only) and its projection in a visual

space (which is said to be *grounded* since visual information has been incorporated) have complementary information, that can be exploited by concatenating them. Thus, we also evaluate various concatenation combinations between X and its grounded projections.

Evaluations are reported in Table 4.5 for standard word embeddings benchmarks, as done in (Collell et al. 2017) we use WordSim353 (Finkelstein et al. 2002), MEN (Bruni et al. 2014), SimLex-999 (F. Hill et al. 2015), SemSim and VisSim (Silberer et al. 2014), along with the concreteness prediction task on the USF dataset (Nelson et al. 2004). We observe that grounded vectors (i.e. $V_{\text{trans}}(X)$ and $V_{\text{sup}}(X)$) tend to outperform X on all benchmarks except Concreteness, and that the information present in X and the grounded vectors is complementary, as $X \oplus V_{\text{trans}}(X)$ (resp. $X \oplus V_{\text{sup}}(X)$) outperforms both X and $V_{\text{trans}}(X)$ (resp. $V_{\text{sup}}(X)$). Interestingly, we notice that $X \oplus V_{\text{trans}}(X)$ gives the highest performance, thus showing the efficiency of exploiting unsupervised visual information using CM-GAN.

4.4.3 Zero-Shot Sentence-to-Image Matching

	Text to Image				**Image to Text**			
Model	MFR	R_1	R_5	R_{10}	MFR	R_1	R_5	R_{10}
Rand.	50	0.1	0.5	1	50	0.1	0.5	1
CONSE	20.3	3.3	13.9	23.5	14.1	2.8	12.3	20.2
$\mathcal{L}_{gan}$	21	4.2	15.7	25.2	14.1	3.1	15.1	23.7
$\mathcal{L}_{c}$	18.3	4.7	14.3	24	11.3	3.7	**16.6**	27.3
$\mathcal{L}_{cgan}$	**15.7**	**5.7**	**17.2**	**27.7**	**11.1**	**4.1**	15.5	**27.7**
$\mathcal{L}_{sup}$	*7.3*	*8.3*	*26.3*	*39.8*	*4.5*	*7.1*	*28.1*	*44.7*

Table 4.6 – Cross-modal retrieval results on MS COCO.

In this section, we demonstrate that CycleGan can align the visual and textual modalities *without any supervision*. We also show that CM-GAN can be applied to other settings than ImageNet: (i) replacing words by sentences, and (ii) with as many classes as training examples (590K sentences).

We evaluate several scenarios of our model on the standard cross-modal retrieval task: given a sentence, retrieving the closest image (Text to Image) and vice versa (Image to Text settings) — see Section 2.4.4 of the Background Chapter for an overview of Cross-Modal Retrieval — in Table 4.6. We observe that:

- the initialization (CONSE) already shows substantial improvement compared to the Random baseline,
- models based on CycleGAN improve performances compared to the CONSE model, due to the beneficial action of the CycleGAN loss,

- CM-GAN has much lower performance than the Supervised baseline (where $\mathcal{L}_{sup}$ is learned), which is intuitively expected.

4.5 Conclusion

4.5.1 Summary of the contributions

In this Chapter, we propose a new model for Transductive ZSL, and evaluate it on ImageNet, which has a high number of unseen classes (20K). Our model relies on a supervised loss to align seen classes data, and a transductive CycleGAN loss to align unseen classes data. We demonstrate that:

- CM-GAN is very efficient on the ImageNet T-ZSL task, with state-of-the-art results,
- visual and textual modalities can be somewhat aligned without supervision on a zero-shot sentence-to-image task on MS COCO,
- textual representations can be enhanced using CM-GAN.

We thus provide the following answer to the Chapter Question:

- **CQ**: There is meaningful information to be exploited in the similarities between textual and visual distributions, even in cases where there is no direct text/vision supervision. We showed it for three multimodal tasks: Transductive Zero-Shot Learning, Zero-Shot Image-to-Sentence Matching and Visual Grounding of Language.

4.5.2 Perspectives

In this section, we present research perspectives following the contribution made in this Chapter.

Zero-shot sentence-to-image matching In the present Chapter, we performed zero-shot sentence-to-image matching by training a Cross-Modal CycleGAN model on MS COCO data. We represented a sentence by the sum of its word embeddings, and we represented images using a convex combination of the most probable class labels embeddings, with the CONSE model (Norouzi et al. 2014). We showed that the CycleGAN model could, without any supervision, capture some text/vision alignment; however, improvements compared to the CONSE initialization could be done with a new method to encode sentence, able to take into account word order.

An interesting perspective would be to encode sentences and images using BERT. To encode images using BERT, we can use the model that we propose in Chapter 5, in which an image is seen as a sequence of objects, and object features are projected into BERT's embedding layer using a cross-modal linear layer. The only learnable parameters would be the latter linear projection layer. Chapter 5 gives us some hints that this method may show good results, as BERT contain abstractions that generalize across modalities.

Zero-Shot Learning from Noisy Text Description Even though ZSL is mostly focused on classifying images with class labels (e.g, *Abyssinian cat*, *dog*, *traffic light*), some works (Elhoseiny et al. 2017; Yizhe Zhu et al. 2018) have proposed, instead, to consider *noisy* text descriptions e.g, *The Parakeet Auklet is a small (23cm) auk with a short orange bill that is upturned ...* We could apply our CrossModal CycleGAN model to this task: to do so, we would need to select useful information in the text (visual words like *small* or *orange* for example) and represent the text as a weighted linear combination of these salient words. We could even consider a fully unsupervised configuration (as done in the zero-shot text-to-image matching task) in which we do not use a direct supervision between images and textual descriptions.

CHAPTER 5

ON THE CROSS-MODAL TRANSFERABILITY OF LANGUAGE MODELS

Contents

Chapter abstract

Pre-trained language models such as BERT have recently contributed to significant advances in Natural Language Processing tasks. Interestingly, while multilingual BERT models have demonstrated impressive results, recent works have shown how monolingual BERT can also be competitive in zero-shot cross-lingual settings. This suggests that the abstractions learned by these models can transfer across languages, even when trained on monolingual data. In this paper, we investigate whether such generalization potential applies to other modalities, such as vision: does BERT contain abstractions that generalize beyond text? *We introduce* BERT-gen, *an architecture for text*

generation based on BERT, able to leverage on either mono- or multi- modal representations. The results reported under different configurations indicate a positive answer to our research question, and the proposed model obtains substantial improvements over the state-of-the-art on two established Visual Question Generation datasets.

This work is currently under review at the EMNLP 2020 conference:

• *Thomas Scialom*, Patrick Bordes*, Paul-Alexis Dray, Jacopo Staiano, Patrick Gallinari "What BERT sees: Cross-Modal Transfer for Visual Question Generation".*

5.1 Introduction

5.1.1 Positioning

In the present Chapter, we explore the cross-modal capabilities of language models, by applying a state-of-the-art language model — BERT (Devlin et al. 2019) — to a multimodal task: Visual Question Generation. This contribution is linked to the following research questions:

- **RQ1** (*Can vision help to refine language understanding ?*): Visual Question Generation is a multimodal extension of the Question Generation task: it is thus a *NLP task aided by CV*. One of the challenge is to determine whether the visual modality enables to generate more salient questions compared to a textual input. Studying the impact of vision can be done (i) quantitatively using ablation studies and (ii) qualitatively using attention visualizations.

- **RQ2** (*Can language help to refine visual understanding ?*): The main question in this Chapter is whether the BERT model, which is trained on textual data, possesses abstractions that generalize to the visual modality. To test this hypothesis, we propose to integrate visual data within BERT without pre-training, using a simple linear layer, and to evaluate the resulting Multimodal BERT model on the VQG task.

- **RQ3** (*Can modalities be translated into one another?*): In VQG, the goal is to generate a meaningful question from a multimodal input. A question is different than a caption in the sense that it does not summarize the high-level content of an image, but rather extends its content, or tries to clarify some grey areas. To understand the contribution of each modality regarding the quality of the generated question, we evaluate three versions of our model: (i) textual input only, (ii) visual input only and (iii) textual and visual input simultaneously.

As a result, we derive Chapter Questions (CQ) that we strive to answer throughout this Chapter:

- **CQ1**: Does the visual modality help to generate more meaningful question in the Question Generation task?
- **CQ2**: Does BERT contain abstractions that generalize to the visual modality?
- **CQ3**: What is the impact of each modality when generating a question?

5.1.2 Visual Question Generation

In the Artificial Intelligence community, several works have investigated the longstanding research question of whether textual representations encode visual information. On the one hand, a large body of research called *language grounding* considers that textual representations lack visual commonsense (Baroni 2016), and intend to *ground* the meaning of words (A. Lazaridou et al. 2015a; G. Collell et al. 2017) and sentences (Kiela et al. 2018; P. Bordes et al. 2019) in the perceptual world. In another body of work, textual representations have successfully been used to tackle multi-modal tasks (Baltrusaitis et al. 2019) such as Zero-Shot Learning (E. Zablocki et al. 2019), Visual Question Answering (Malinowski et al. 2014) or Image Captioning (Socher et al. 2014b). Following the latter line of research, in this paper we evaluate the potential of pre-trained language models to generalize in the context of Visual Question Generation (VQG) (Mostafazadeh et al. 2016).

The Visual Question Generation task — that we present in Section 2.2.2.1 of the Background Chapter — allows us to investigate the cross-modal capabilities of BERT: unlike Image Captioning (where the input is only visual) or VQA (where the input is visual *and* textual), VQG is a multi-modal task where input can be textual *and/or* visual. VQG data usually includes images and the associated captions, along with corresponding questions about the image; thus, different experimental setups can be designed to analyze the impact of each modality. Indeed, the questions can be generated using *i)* textual (the caption), *ii)* visual (the image), or *iii)* multi-modal (both the caption and the image) input.

From a practical standpoint, the VQG task has several applications: robots or AI assistants could ask questions rooted in multi-modal data (e.g. fusing conversational data with visual information from captors and cameras), in order to refine their interpretation of the situation they are presented with. It could also allow systems relying on knowledge-bases to gain visual common sense and deal with the Human Reporting Bias (Misra et al. 2016), which states that the content of images and text are intrinsically different, since visual common sense is rarely explicitly stated in text.

5.1.3 Mono- and Multi-modal Neural Language Models

The BERT language model (Devlin et al. 2019) is a Deep Bidirectional Transformer (Vaswani et al. 2017) pre-trained on textual corpora (BookCorpus and Wikipedia) using a Masked Language Model (MLM) objective – predicting some words that are randomly masked in the sentence, along with a sentence entailment loss. Recent research efforts (Artetxe et al. 2019) have shown how BERT encodes abstractions that generalize across languages, even when trained on monolingual data only. This contradicts the common belief (Pires et al. 2019; S. Wu et al. 2019) that a shared vocabulary and joint training on multiple languages are essential to achieve cross-lingual generalization capabilities. In this work, we further investigate the generalization potentials of large pre-trained LMs, this time moving to a cross-modal setup: *does BERT contain abstractions that generalize beyond text?*

Recently, BERT-based Multi-Modal Language Models have been proposed (J. Lu et al. 2019; Tan et al. 2019; L. H. Li et al. 2019; W. Su et al. 2019) to tackle multi-modal tasks, using different approaches to incorporate visual data within BERT. Some works use *single-stream* Transformers in which visual features are incorporated in a BERT-like Transformer, e.g, VisualBERT (L. H. Li et al. 2019), while others employ modality-specific encoders built on standard Transformer blocks, which are then fused into a cross-modal encoder, such as ViLBERT (J. Lu et al. 2019).

5.1.4 Contributions

From the literature on Multimodal Neural Language Models, it is left to explore whether the cross-modal alignment is fully learned, or it is to some extent already encoded in the BERT abstractions. Therefore, in contrast with those approaches, we explicitly avoid using the following complex mechanisms:

- *Multi-modal supervision*: all previous works exploit an explicit multi-modal supervision through a pre-training step; the models have access to text/image pairs as input, to align their representations. In contrast, our model can switch from text-only to image-only mode without any explicit alignment.

- *Image-specific losses*: specific losses such as Masked RoI (Region of Interest) Classification with Linguistic Clues (W. Su et al. 2019) or sentence-image prediction (L. H. Li et al. 2019) have been reported helpful to align visual and text modalities. Instead, we only use the original MLM loss from BERT (and not its entailment loss).

- *Non-linearities*: we explore a scenario in which the only learnable parameters, for aligning image representations to BERT, are those of simple linear projection

layer. This allows us to assess whether the representations encoded in BERT can transfer *out-of-the-box* to another modality.

Furthermore, to the best of our knowledge, this paper is the first attempt to investigate multi-modal text *generation* using pre-trained language models. We introduce *BERT-gen*, a text generator based on BERT, that can be applied both in mono and multi-modal settings. We treat images similarly to text: while a sentence is seen as a sequence of (sub)word tokens, an image is seen as a sequence of objects associated to their corresponding positions (bounding boxes). We show how a simple linear mapping, projecting visual embeddings into the first layer, is enough to ground BERT in the visual realm: text and image object representations are found to be effectively aligned, and the attention over words transfers to attention over the relevant objects in the image.

Our contributions can be summarized as follows:

1. we introduce *BERT-gen*, a novel method for generating text using BERT, that can be applied in both mono and multi-modal settings;
2. we show that the semantic abstractions encoded in pre-trained BERT can generalize to another modality;
3. we report state-of-the art results on the VQG task;
4. we provide extensive ablation analyses to interpret the behavior of *BERT-gen* under different configurations (mono- or multi- modal).

5.2 Model

In VQG, the objective is to generate a relevant question from an image and/or its caption. The caption X_{txt} is composed of M tokens $txt_1, ..., txt_M$; these tokens can be words or subwords (smaller than word) units depending on the tokenization strategy used. As BERT uses subword tokenization, throughout this paper we will refer to subwords as our tokenization units.

The proposed model is illustrated in Figure 5.1. In Section 5.2.1, we detail how images are incorporated in the Transformer framework. In Section 5.2.2, we present *BERT-gen*, a novel approach to use BERT for text generation.

5.2.1 Representing an Image as Text

In this work, we treat textual and visual inputs similarly, by considering both as sequences. Since an image is not a priori sequential, we consider the image X_{img} as a sequence of object regions $img_1, ..., img_N$, as described below.

The images are first processed as in Tan et al. 2019: a Faster-RCNN (S. Ren et al. 2017), pre-trained on Visual Genome (Krishna et al. 2017), detects the

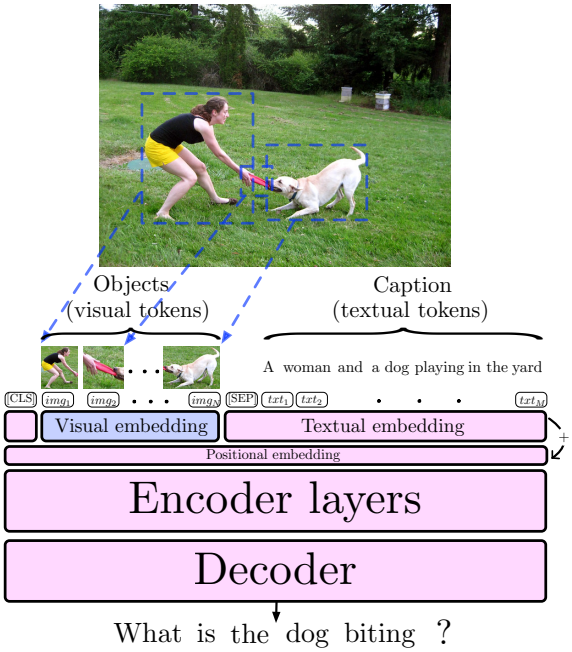

Figure 5.1 – Model overview. Captions are encoded via BERT embeddings, while visual embeddings (blue) are obtained via a linear layer, used to project image representations to the embedding layer dimensions.

$N = 36$ most salient regions (those likely to include an object) per image. The weights of the Faster-RCNN are fixed during training, as we use the precomputed representations made publicly available [1] by Anderson et al. 2018. Each image is thus represented by a sequence of $N = 36$ semantic embeddings $f_1, ...f_N$ (one for each object region) of dimension 2048, along with the corresponding bounding box coordinates $b_1, ...b_N$ of dimension 4. With this approach, the BERT attention can be computed at the level of objects or salient image regions; had we represented images with traditional CNN features, the attention would instead correspond to a uniform grid of image regions without particular semantics, as noted in Anderson et al. 2018. To build an object embedding o_j encoding both the object region semantics and its location in the image, we concatenate f_j and b_j ($j \in [1, N]$). Hence, an image is seen as a sequence of $N = 36$ visual representations (each corresponding to an object region) $o_1, ..., o_N$. Object region representations o_i are ordered by the prediction certainty of the object detected, and the model has access to their relative location in the image through the vectors b_i.

To investigate whether our BERT-based model can transfer knowledge beyond language, we consider image features as simple visual tokens that can be presented to the model *analogously* to textual token embeddings. In order to make the o_j vectors (of dimension $2048 + 4 = 2052$) comparable to BERT embeddings (of dimension 768), we use a simple linear *cross-modal projection* layer W of dimensions

1. https://github.com/peteanderson80/bottom-up-attention

2052×768. The N object regions detected in an image, are thus represented as $X_{img} = (W.o_1, ..., W.o_N)$. Once mapped into the BERT embedding space with W, the image is seen by the rest of the model as a sequence of units with no explicit indication if it is a text or an image embedding.

5.2.2 *BERT-gen*: Text Generation with BERT

We cast the VQG task as a classic sequence-to-sequence (Sutskever et al. 2014) modeling framework:

$$P_{\Theta,W}(Y|X) = \prod_{t=1}^{T} P_{\Theta,W}(y_t|X, y_{<t}) \tag{5.1}$$

where the input $X = X_{txt}$ in caption-only mode, $X = X_{img}$ in image-only mode, and $X = X_{img} \oplus X_{txt}$ in a multi-modal setup; $Y = y_1, ..., y_T$ is the question composed of T tokens. Θ are the parameters of the BERT model [2]; W represents the weights of the linear layer used for projecting visual input to the BERT embedding layer.

As mentioned earlier, BERT is a Transformer (Vaswani et al. 2017) encoder pre-trained using the Masked Language Model (MLM) objective: tokens within the text are replaced with a `[MASK]` special token, and the model is trained to predict them. Since BERT was not trained with an unidirectional objective, its usage for text generation is not straightforward.

To generate text, Yang Liu et al. 2019 propose to stack a Transformer decoder, symmetric to BERT. However, the authors report training difficulties since the stacked decoder is not pre-trained, and propose a specific training regime, with the side-effect of doubling the number of parameters. Dong et al. 2019 opt for an intermediate step of self-supervised training, introducing a unidirectional loss. As detailed below, we propose a relatively simpler, yet effective, method to use BERT *out-of-the-box* for text generation.

Decoder We simply use the original BERT decoder as is, initially trained to generate the tokens masked during its pre-training phase. It consists in a feed-forward layer, followed by normalization, transposition of the embedding layer, and a softmax over the vocabulary.

Next Token Prediction At inference time, to generate the first token of the question y_1, we concatenate `[MASK]` to the input tokens X, then encode $X \oplus$ `[MASK]` with the BERT encoder, and feed the output of the encoder to the decoder; y_1 is the output of the decoder for the `[MASK]` token. Subsequently, given y_1, we

2. We use the smaller architecture released, `BERT-base` (12 layers), pre-trained on English cased text.

concatenate it to the input tokens and encode $X \oplus y_1 \oplus$ [MASK] to predict the next token y_2. This procedure is repeated until the generation of a special token [EOS] signaling the end of the sentence.

Attention Trick As we iteratively concatenate the generated tokens, the BERT omni-directional self-attention mechanism would impact, at every new token, the representations of the previous tokens. To counter that, we use a *left-to-right* attention mask, similar to the one employed in the original Transformer decoder (Vaswani et al. 2017). For the input tokens in X, we apply such mask to all the target tokens Y that were concatenated to X, so that input tokens can only attend to the other input tokens. Conversely, for target tokens y_t, we put an attention mask on all tokens $y_{>t}$, allowing target tokens y_t to attend only to the input tokens and the already generated target tokens.

This novel method allows to use pre-trained encoders for text generation. In this work, we initialize our model with the parameters from BERT-base. Nonetheless, the methodology can be applied to any pre-trained Transformer encoders such as RoBERTa (Yinhan Liu et al. 2019), or Ernie (Y. Sun et al. 2019).

Modality-specific setups The proposed model can be used in either mono- or multi- modal setups. This is accomplished by activating the textual and/or visual modules.

5.3 Experimental Protocol

Our main objective is to measure whether the textual knowledge encoded in pre-trained BERT can be beneficial in a cross-modal task. Thus, we define the three experimental setups that follow from each other, which we refer to as Step 1, 2, and 3:

1. Caption only Deactivating the *Visual embedding* (see Figure 5.1), the model has only access to textual input, *i.e.* the caption. The model is initialized with the BERT weights and trained according to Equation 5.1.

2. Image only Conversely, deactivating the *Textual embedding* module (see Figure 5.1), the model has only access to the input image, not the caption. To indicate the position t of img_t in the sequence, we sum the BERT positional embedding of t to the visual representation of img_t, just as we would do for a text token txt_t. The model is initialized with the weights learned during step 1. All *BERT-gen* Θ weights are frozen, and only the linear layer W is learnable. Hence, *if the model is able to learn to generate contextualized questions w.r.t. the image, it shows that a simple linear layer is enough to bridge the two modalities*.

3. Image + Caption The full model is given access to both image and caption inputs. In this setup, we separate the two different inputs by a special BERT token `[SEP]`. Thus, the input sequence for the model takes the form of $\texttt{[CLS]}, img_1, ..., img_N, \texttt{[SEP]}, txt_1, ..., txt_M$.

In step 1, only *BERT-gen* Θ parameters are learned, as no image input was given. In step 2, W is trained while keeping Θ frozen. Finally then, in step 3, we fine-tune the model using both image and text inputs: the model is initialized with the parameters Θ learned during step 1 and the W learned during step 2, and we unfreeze all parameters.

Ablations Additionally, we report results obtained with: *Image only (unfreeze)*, where the *BERT-gen* parameters Θ are not frozen; and *Image+Caption (from scratch)* where the model is learned without the intermediate steps 1 and 2: the *BERT-gen* parameters Θ are initialized with the weights from pre-trained BERT while W is randomly initialized.

5.3.1 Datasets

We conduct our experiments using two established datasets for Visual Question Generation:

VQG_{COCO} Introduced by Mostafazadeh et al. 2016, it contains 2500 training images, 1250 validation images and 1250 test images from MS COCO (T. Lin et al. 2014b); each image has 5 corresponding questions and 5 ground-truth captions.[3]

VQA The Visual Question Answering (Teney et al. 2016) dataset can be used to derive VQG data (Y. Li et al. 2018). The task is reversed: instead of answering the question based on the image (VQA), models are called to generate a relevant question given the image (VQG). Also based on MS COCO, it contains 82783 training images, 40504 validation images and 81434 testing images. In $VQA1.0$,[4] each image has 3 associated questions. Since the test set of MS COCO does not contain ground-truth captions, we generated artificial captions for it using NeuralTalk2 (Karpathy et al. 2017), which is a standard image captioning model: for fair comparison, we used exactly the same model[5] as Patro et al. 2019 (MDN-Joint).

3. Publicly available at https://www.microsoft.com/en-us/download/details.aspx?id=53670
4. Publicly available at https://visualqa.org/vqa_v1_download.html
5. Publicly available at https://github.com/karpathy/neuraltalk2

5.3.2 Baselines

We compare the proposed model to the following:

Sample (Y. Yang et al. 2015) Questions are generated by a RNN conditioned on the image: at each generation step, the distribution over the vocabulary is computed and used to sample the next generated word. This baseline enables to generate diverse questions over the same image, as the word selection process is non-deterministic.

Max (Y. Yang et al. 2015) Using the above model, selecting words with maximum probability from the computed distribution.

MDN-Joint (Patro et al. 2019) State-of-the-art model on $VQA1.0$, based on joint usage of caption and image information.

MC-SBN (Patro et al. 2020) State-of-the-art on VQG_{COCO}. The model jointly leverages on multiple cues (the image, place information, caption, tags) to generate questions.

5.3.3 Metrics

We report the following metrics for all experiments, consistently with previous works:

BiLingual Evaluation Understudy (BLEU) (Papineni et al. 2002) A precision-oriented metric, originally proposed to evaluate machine translation. It is based on the counts of overlapping n-grams between the generated sequences and the human references.

Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (C.-Y. Lin 2004) The recall-oriented counterpart to BLEU metrics, again based on n-gram overlaps.

Metric for Evaluation of Translation with Explicit ORdering (METEOR) (Banerjee et al. 2005) The harmonic mean between precision and recall w.r.t. unigrams. As opposed to the other metrics, it also accounts for stemming and synonymy matching.

Consensus-based Image Description Evaluation (CIDEr) (Vedantam et al. 2015a) Originally designed for Image Captioning, it uses human consensus among the multiple references, favoring rare words and penalizing frequent words. This feature is particularly relevant for our task, as the automatically generated questions

often follow similar patterns such as "What is the [...] ?". Indeed, we verify experimentally (cf Table 5.1 and Table 5.2) that the CIDEr metric is the most discriminant in our quantitative results.

5.3.4 Implementation details

All models are implemented in PyText (Aly et al. 2018). For all our experiments we used a single NVIDIA RTX 2080 Ti GPU, a batch size of 128 and 5 epochs. We used the Adam optimizer with the recommended parameters for BERT: learning rate is set at $2e^{-5}$ with a warmup of 0.1. The most computationally expensive experiment is the step 3 described above: for this model, completion of one epoch demands 30 seconds and 2 minutes for VQG_{COCO} and VQA datasets, respectively. Metrics were computed using the Python package released by Du et al. 2017. [6]

5.4 Results

Model	Step	BLEU1	BLEU2	BLEU3	BLEU4	ROUGE-L	METEOR	CIDEr
Sample		38.8	-	-	-	34.2	12.7	13.3
Max		59.4	-	-	-	49.3	17.8	33.1
MDN-Joint		65.1	-	-	-	52.0	22.7	33.1
Caption	1	75.41	56.49	43.26	32.28	66.18	26.51	43.56
Image	2 (freeze)	73.62	53.54	39.37	27.44	64.34	24.36	29.58
Image	2 (unfreeze)	73.97	55.07	42.20	31.76	65.70	26.36	41.43
Image+Cap.	3	75.59	**56.88**	**43.96**	**33.35**	**66.71**	**26.76**	**44.99**
Image+Cap.	3 (f. scratch)	**75.84**	56.42	43.53	32.85	66.30	25.92	38.81

Table 5.1 – Quantitative VQG results on $VQA1.0$. We report results from previous works in the upper block, and those obtained by our proposed models in the bottom block.

In Table 5.1, we report quantitative results for the VQG task on $VQA1.0$ (where captions are automatically generated). The *Caption only* model already shows strong improvements for all metrics over state-of-the-art models. For this text-only model, the impressive performance can mostly be attributed to BERT, demonstrating once again the benefits obtained using pre-trained language models. In our second step (*Image only*), the BERT Θ parameters are frozen and only those of the cross-modal projection matrix W are learned. Despite using a simple linear

6. https://github.com/xinyadu/nqg/tree/master/qgevalcap

layer, the model is found to perform well, generating relevant questions given only visual inputs.

This suggests that the conceptual representations encoded in pre-trained language models such as BERT can effectively be used beyond text. Further, we report an additional *Image only* experiment, this time unfreezing the BERT parameters Θ – see *Step 2 (unfreeze)* in Table 5.1. As could be expected, since the model is allowed more flexibility, the performance is found to further improve.

Finally, in our third step (*Image + Caption*), we obtain the highest scores, for all metrics. This indicates that the model is able to effectively leverage the combination of textual and visual inputs. Indeed, complementary information from both modalities can be exploited by the self-attention mechanism, making visual and textual tokens interact to generate the output sequences. Again, we additionally report the results obtained bypassing the intermediate steps 1 and 2: for the model denoted as *Step 3 (from scratch)* (last row of Table 5.1), Θ parameters are initialized with the original weights from pre-trained BERT, while the W matrix is randomly initialized. Under this experimental condition, we observe lower performances, a finding that consolidates the importance of the multi-step training procedure we adopted.

Model	Step	BLEU1	BLEU2	BLEU3	BLEU4	ROUGE-L	METEOR	CIDEr
MDN-Joint		36.0	24.9	16.8	10.4	41.8	23.4	50.7
MC-SBN		40.7	-	-	-	-	22.6	-
Caption	1	**74.58**	54.94	43.33	34.36	64.09	**29.76**	**77.70**
Image	2 (freeze)	69.57	49.93	38.23	29.54	61.01	27.03	57.38
Image	2 (unfreeze)	74.34	**55.26**	**43.47**	**34.41**	**64.63**	29.17	72.18
Image+Cap.	3	70.96	50.83	39.20	30.29	61.87	27.65	62.77
Image+Cap.	3 (f. scratch)	64.18	42.88	30.19	20.14	56.99	23.32	30.99
Human		*86*	-	-	-	-	*60*	-

Table 5.2 – Quantitative VQG results on VQG_{COCO}. We report results from previous works in the upper block, and those obtained by the our proposed models in the middle block. Human Performance is taken from Mostafazadeh et al. 2016.

In Table 5.2, we report quantitative VQG results on VQG_{COCO}. These are globally consistent with the ones above for $VQA1.0$. However, we observe two main differences. First, a bigger relative improvement over the state-of-the-art. As the efficacy of pre-trained models is boosted in small-data scenarios (Radford et al. 2018), this difference can be explained by the smaller size of VQG_{COCO}. Second, we note that the *Caption only* model globally outperforms all other models, especially on the discriminant CIDEr metric. This can be explained by

	Read.	Caption Rel.	Image Rel.
Caption only	**4.9**	**4.72***	4.25*
Image only	4.77	3.87	4.32*
Image + Caption	4.89	4.06*	**4.69***
Human	*4.83*	*3.64*	*4.9*

Table 5.3 – Human evaluation results for three criterions: *readability*, *caption relevance* and *image relevance*. Two-tailed t-test results are reported in comparison to "Human" (*: $p < 0.05$).

the fact that, in VQG_{COCO}, the captions are human-written (whereas they are automatically generated for $VQA1.0$) and, thus, of higher quality; moreover, the smaller size of the dataset could play a role hindering the ability to adapt to the visual modality. Nonetheless, the strong performances obtained for *Step 2* compared to the baselines highlight the effectiveness of our method to learn a cross-modal projection even with a relatively small number of training images. Finally, the model *Image+Caption* (from sratch) obtains the worst performances, showing the interest of proceeding to a step-by-step learning process, especially for the CIDEr metric (30.99 vs 77.70 for Step 1).

Human Evaluation To get more in-depth understanding of our models, we report human assessment results in Table 5.3. We randomly sampled 50 images from the test set of $VQA1.0$. Each image is paired with its caption, the human-written question used as ground-truth, and the output for our three models: *Caption only*, *Image only* and *Image+Caption*. We asked 3 human annotators to assess the quality of each question using a Likert scale ranging from 1 to 5, for the following criteria: *readability*, measuring how well-written the question is; *caption relevance*, how relevant the question is w.r.t. to the caption; and, *image relevance*, how relevant the question is toward the image. For caption and image relevance, the annotators were presented with only the caption and only the image, respectively.

We observe that all evaluated models produce well-written sentences, as *readability* does not significantly differ compared to human's questions. Unsurprisingly, the *Caption only* model shows a higher score for *caption relevance*, while the relatively lower *image relevance* score can be explained by the automatically generated and thus imperfect captions in the $VQA1.0$ dataset. Comparatively, the *Image only* model obtains lower *caption relevance* and higher *image relevance* scores; this indicates that the cross modal projection is sufficient to bridge modalities, allowing BERT to generate relevant questions toward the image. Finally, the *Image + Caption* model obtains the best *image relevance* among our models, consistently the quantitative results reported in Table 5.1 and Table 5.2.

5.5 Model Discussion

What does the model look at? To interpret the behavior of attention-based models, it is useful to look at which tokens are given higher attention (Clark et al. 2019). In Figure 5.2, we present two images *A* and *B*, along with their captions and the three generated questions corresponding to our three experimental setups (*Caption only*, *Image only* and *Image + Caption*). For this analysis, we average the attention vectors of all the heads in the last layer, and highlight the textual and visual tokens most attended by the models.

For both images, the *Caption only* model attends to salient words in the caption. The *Image only* model remains at least as much relevant: on image *A*, it generates a question about a table (with an unclear attention). Interestingly, for image *B*, the *Image only* model corrects a mistake from step 1: it is a *woman* holding an umbrella rather than a *man*, and the attention is indeed focused on the woman in the image. Finally, the *Image + Caption* model is able to generate fitting questions about the image, with relatively little relevance to the caption: for image *A*, *Image + Caption* the model generates "What time is it?" while paying attention to the clock; for image *B*, *Image + Caption* generates "What is the color of the umbrella ?", focusing the attention on the umbrella. The captions of either samples include no mentions of clocks or umbrellas, further indicating effective alignment between visual and textual representations.

Cross-modal alignment We carry out an additional experiment to analyze the text/vision alignment for each model. Figure 5.3 shows the *cross-modal* similarity X_{sim} for different model scenarios, computed at each BERT-base layer from 1 to 12. We define the cross-modal similarity X_{sim} as the cosine similarity between the vector representations of both modalities. For all models (*Random Transformer*, *Caption only*, *Image only*, *Image+Caption*), these vectors are the two continuous space representations from a model when given as input either *i)* an image, or *ii)* its corresponding caption. Please note that, for *Random Transformer* and *Caption only*, visual embeddings are computed with a matrix W that is random. We represent these captions and images vectors with the special BERT token `[CLS]`, following previous works (Reif et al. 2019) where `[CLS]` is used to represent the entire sequence.

The reported values correspond to the average cross-modal similarity calculated for all the examples of VQG_{COCO} test set. In addition to the setups described in Section 5.3 (*Caption-only*, *Image-only* and *Image + Caption*), we also report X_{sim} for *Random Transformer*, a BERT architecture with random weights. As expected, its X_{sim} is close to zero.

All the other models are based on BERT. As suggested by Tenney et al. 2019, the first layers in BERT tend to encode lower-level language information. This might explain why the models have similar X_{sim} scores up to the 9th layer, and

A

A room with a desk and a laptop

(1) What is the color of the desk ?

(2) What is the color of the table ?

(3) What time is it ?

B

A group of people standing on a street

(1) What is the man holding ?

(2) What is the woman holding ?

(3) What is the color of the umbrella ?

Figure 5.2 – Qualitative Analysis. We show the outputs of the three steps of our model, using two samples from the *VQA*1.0 test set.
1) Caption only; *2)* Image only; *3)* Image + Caption. Words and object regions with maximum attention are underlined and marked, respectively. Color intensity is proportional to attention.

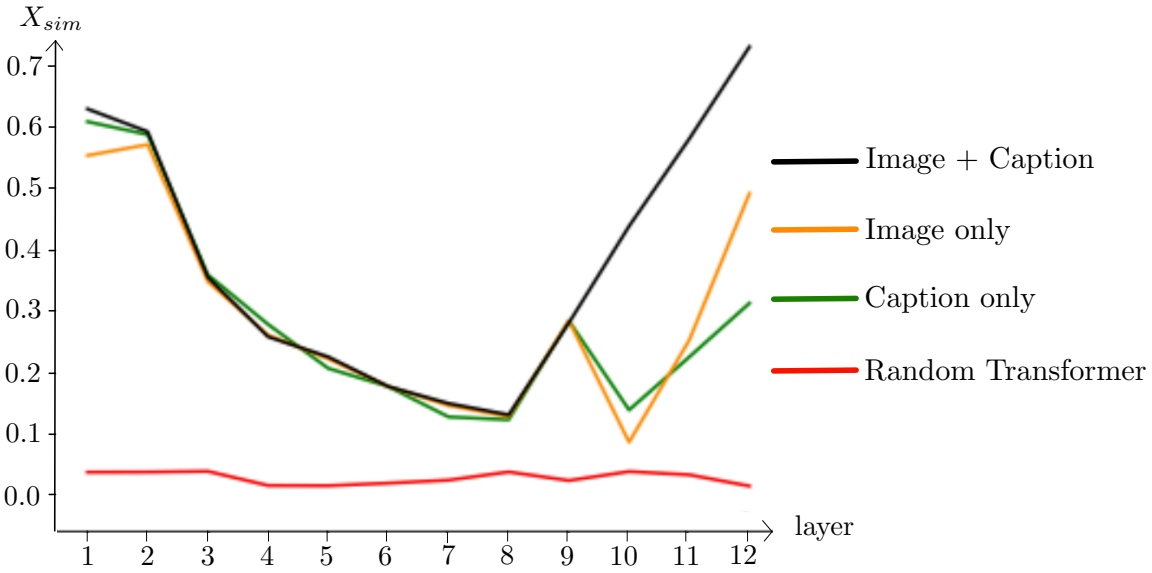

Figure 5.3 – Cross-modal similarity X_{sim} between images in VQG_{COCO} and corresponding captions at each BERT encoding layer. Captions and images are embedded here using the `[CLS]` special token.

diverge afterwards: the weights for those layers remain very similar between our fine-tuned models.

For the last layer ($l = 12$), we observe that *Caption only* $<$ *Image only* $<$ *Image + Caption*. The interpretation is the following. The *Caption only* model has never seen images during training, and therefore is not able to encode meaningful semantic information of images. Still, its reported $X_{sim} > 0$ can be attributed to the fact that, when fine-tuned on the Visual Question Generation (VQG) task during Step 1, *BERT-gen* encodes task-specific information in the `[CLS]` token embedding (*e.g.* a question ends with a "?" and often begins with "What/Where/Who"). *Image only* $>$ *Caption only* can be explained by the learning of the cross-modal projection W. However, since BERT is not fine-tuned, the model learns a "contortion" allowing it to align text and vision. Finally, *Image + Caption* $>$ *Image only* can be attributed to BERT fine-tuning, contributing to an increase in the observed gap, and its emergence in earlier layers.

5.6 Conclusion

5.6.1 Summary of the contributions

We investigated whether the abstractions encoded in a pre-trained BERT model can generalize beyond text. We proposed *BERT-gen*, a novel methodology that allows to directly generate text from *out-of-the-box* pre-trained encoders, either in

mono- or multi- modal setups. Moreover, we applied *BERT-gen* to Visual Question Generation, obtaining state-of-the-art results on two established datasets. We showed how a simple linear projection is sufficient to effectively align visual and textual representations.

We thus provide the following answers to the Chapter Questions:

- **CQ1**: We showed that the visual modality enables to generate better questions: thanks to the attention module, our model can focus on relevant elements of the image and ask meaningful questions.
- **CQ2**: BERT contains abstractions that generalize beyond text, as a simple linear layer is sufficient for reaching substantial results on VQG.
- **CQ3**: We showed, using human evaluation, that using both modalities simultaneously as input enables to generate questions that are highly relevant toward the image. Similarly, quantitative results highlight that the text+vision input setting reaches highest performances.

5.6.2 Perspectives

In this section, we present research perspectives following the contribution made in this Chapter.

Apply *BERTgen* to other generative multimodal tasks In this Chapter, we extended the BERT model to generate sequential output, and apply it to the VQG task. We derive three configurations: caption-to-question, image-to-question and image+caption-to-question. Other multimodal tasks could be tackled with *BERTgen*.

For the Image Captioning task — see Section 2.4.1 of the Background Chapter for an overview — we could first train *BERTgen* with a caption-to-caption task (train the model to predict a synonym caption). Then, we could train *BERTgen* in an image-to-caption configuration, by just learning a linear layer to project images into BERT embedding layer. Finally, we could fine-tune our model by training on caption+image-to-caption. Obtaining *BERTgen*'s results for Image Captioning could enable us to: (i) further demonstrate the cross-modal capabilities of BERT, and (ii) compare to state-of-the-art captioning models, that do not rely on BERT.

Does BERT's contain abstractions that generalize to other modalities than vision? In this Chapter, we showed that BERT contains abstractions can generalize to the visual modality. Considering other modalities than vision is an interesting research perspective. For example:

- *speech data*: As speech corresponds to natural language, semantics would be close to BERT's abstractions, so we expect *BERTgen* to show substantial cross-modal capabilities.
- *audio data*: For generic (ii) audio data, we could train *BERTgen* to generate a description in natural language (or a question) about an *audio scene* (a recording taken from a real-world location, or from a movie). Here, the recording would be decomposed in a series of shorter audio elements, sequentially fed to *BERTgen*.
- *video data*: we could tackle Video Question Generation (Y. Wang et al. 2019), which is a recent and under-explored task. Instead of considering sequentially objects of the image as done in this Chapter, we could consider the video as a sequence of frames, with each frame projected to BERT's embedding layer using a linear projection.

CHAPTER 6

GROUNDING LANGUAGE IN VISION: THE CASE OF SENTENCES

Contents

Chapter abstract

Language grounding is an active field aiming at enriching textual representations with visual information. Generally, textual and visual elements are embedded in the same representation space, which implicitly assumes a one-to-one correspondence between modalities. This hypothesis does not hold when representing words, and becomes problematic when used to learn sentence representations — the focus of this paper — as a visual scene can be described by a wide variety of sentences. To overcome this limitation, we propose to

transfer visual information to textual representations by learning an intermediate representation space: the grounded space. We further propose two new complementary objectives ensuring that (1) sentences associated with the same visual content are close in the grounded space and (2) similarities between related elements are preserved across modalities. We show that this model outperforms the previous state-of-the-art on classification and semantic relatedness tasks.

The work in this Chapter has led to the publication of a conference paper:

• *Patrick Bordes*, Éloi Zablocki*, Laure Soulier, Benjamin Piwowarski, and Patrick Gallinari (2019). "Incorporating Visual Semantics into Sentence Representations within a Grounded Space". In: EMNLP 2019.*

6.1 Introduction

6.1.1 Positioning

In this Chapter, we present a contribution of the Visual Grounding of Language task, for which we provide an overview in the Section 2.2.1 of the Background Chapter. This Chapter is linked to **RQ1**: *Can vision help to refine language understanding?* Indeed, the majority of grounding works focus on learning word representations. In this contribution, we consider a higher level of granularity: sentences. As a result, we derive a Chapter Question (CQ) that we strive to answer throughout this Chapter:

- **CQ**: Can visual semantics be transferred within sentence representations ?

6.1.2 Visual grounding of language

Representing text by vectors that capture meaningful semantics is a long-standing issue in Artificial Intelligence. Distributional Semantic Models (Mikolov et al. 2013b; Peters et al. 2018) are well-known recent efforts in this direction, based on the *distributional hypothesis* (Harris 1954). They rely on large text corpora to learn word embeddings. At another granularity level, having high-quality and general-purpose sentence representations is crucial for all models that encode sentences into semantic vectors, such as the ones used in machine translation (Bahdanau et al. 2014) or relation extraction (H. Wang et al. 2019). Moreover, encoding semantics of sentences is paramount because sentences describe relationships between objects, and thus convey complex and high-level knowledge better than individual words (Norman 1972).

Relying only on text can lead to biased representations and unrealistic predictions such as "*the sky is green*" (Baroni 2016). Besides, it has been shown that

human understanding of language is *grounded* in physical reality and perceptual experience (Fincher-Kiefer 2001). To overcome this limitation, an emerging approach is to *ground* language in the visual world: this consists in leveraging visual information, usually from images, to enrich textual representations. [1]

Leveraging images resulted in improved linguistic representations on intrinsic and downstream tasks (Bruni et al. 2014; Silberer et al. 2014). In most of these approaches, cross-modal projections are learned to incorporate visual semantics in the final representations (Angeliki Lazaridou et al. 2015; Collell et al. 2017; Kiela et al. 2018). These works rely on paired textual and visual data and the hypothesis of a one-to-one correspondence between modalities is implicitly assumed: an image of an object univocally represents a word. However, there is no obvious reason implying that the structure of the two spaces should match. Indeed, Collell et al. 2017 empirically show that cross-modal projection of a source modality does not resemble the target modality in terms of its neighborhood structure. This is especially the case for sentences, where many different sentences can describe a similar image. Therefore, we argue that learning grounded representations with projections to a visual space is particularly inadequate in the case of sentences.

6.1.3 Contributions

To overcome this issue, we propose an alternative approach where the structure of the visual space is *partially transferred* to the textual space. This is done by distinguishing two types of complementary information sources. First, the *cluster information*: the implicit knowledge that sentences associated with the same image refer to the same underlying reality. Second, the *perceptual information*, which is contained within high-level representations of images. These two sources of information aim at transferring the structure of the visual space to the textual space. Besides, to preserve textual semantics and to avoid an over-constrained textual space, we propose to incorporate the visual information to textual representations using an intermediate representation space that we call *grounded space*, on which cluster and perceptual objectives are trained.

Our contributions are the following:

1. we define two complementary objectives to ground the textual space, based on implicit and explicit visual information;
2. we propose to incorporate visual semantics through the means of an intermediate space, within which the objectives are learned;
3. we perform quantitative and qualitative evaluations on several transfer tasks, showing the advantages of our approach with respect to previous grounding methods.

1. In the Computer Vision community, *grounding* can also refer to the task of linking phrases with image regions (Xiao et al. 2017), but this is not the focus of the present Chapter.

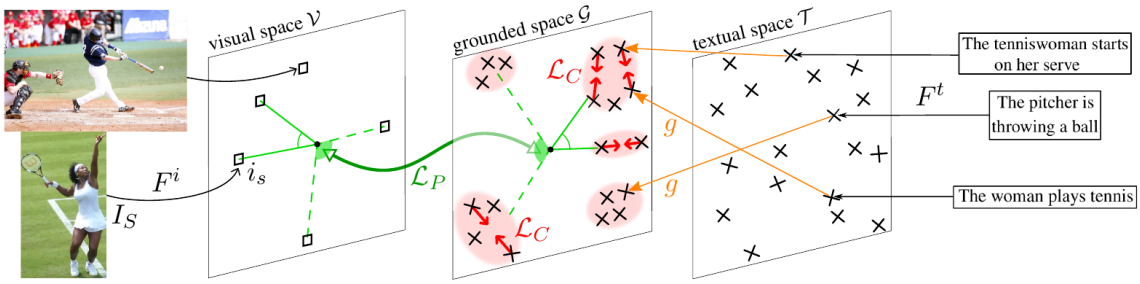

Figure 6.1 – Model overview. Red circles indicate visual clusters. Red arrows represent the gradient of the cluster loss, which gathers visually equivalent sentences — the contrastive term in loss $\mathcal{L}_C$ is not represented. The green arrow and angles illustrate the perceptual loss, ensuring that cosine similarities correlate across modalities. The origin is at the center of each space.

6.2 Incorporating visual semantics within an intermediate grounded space

6.2.1 Model overview

In this work, we aim at learning grounded representations by jointly leveraging the textual and visual contexts of a sentence. We note S a sentence and $s = F^t(S; \theta^t)$ its representation computed with a sentence encoder F^t parametrized by θ^t. We follow the classical approach developed in the language grounding literature at the word level (Angeliki Lazaridou et al. 2015; Eloi Zablocki et al. 2018b), which balances a textual objective $\mathcal{L}_\mathcal{T}$ with an additional grounding objective $\mathcal{L}_\mathcal{G}$:

$$\mathcal{L}(\theta^t, \theta^i) = \mathcal{L}_\mathcal{T}(\theta^t) + \mathcal{L}_\mathcal{G}(\theta^t, \theta^i) \tag{6.1}$$

The parameters θ^t of the sentence encoder F^t are shared in $\mathcal{L}_\mathcal{T}$ and $\mathcal{L}_\mathcal{G}$, and therefore benefit from both textual and grounding objectives. θ^i denotes extra grounding parameters, including the weights of the image encoder F^i. Note that any textual objective $\mathcal{L}_\mathcal{T}$ and sentence encoder F^t can be used. In our experiments, we choose the well-known SkipThought model (Kiros et al. 2015), trained on a corpus of ordered sentences.

In what follows, we focus on the modeling of the grounding objective $\mathcal{L}_\mathcal{G}$, learned on a captioning corpus, where each image is associated with several captions. Grounding approaches generally leverage visual information by embedding textual and visual elements within the same multimodal space (Silberer et al. 2014; Kiela et al. 2018). However, it is not satisfying since texts and images are forced to be in one-to-one correspondence. Moreover, a caption can:

1. have a wide variety of paraphrases and related sentences describing the same scene (e.g., *the kitten is devouring a mouse* versus *a cat eating a mouse*),

2. be visually ambiguous (e.g., *a cat is eating* can be associated with many different images, depending on the visual scene/context),
3. carry non-visual information (e.g., *cats often think about their meals*).

Usual grounding objectives, that embed sentences in the visual space, can discard non-visual information (3) through the projection function. They can handle (1) by projecting related sentences to the same location in the visual space. However, they are over-sensitive to visual ambiguity (2), because ambiguous sentences should be projected to different locations of the visual space, which is not possible with current grounding models.

To overcome this lack of flexibility, we propose the following approach, illustrated in Figure 6.1. To cope with (1), sentences associated with the same image should be close — we call this *cluster information*. To cope with (2), we want to avoid projecting sentences to a particular point of the visual space: instead, we require that the similarity between two images in the visual space (which is linked to the "context discrepancy") should be close to the similarity between their associated sentences in the textual space. We call this *perceptual information*. Finally, as we want to preserve non-visual information in sentence representations (3), we make use of an intermediate space, called *grounded space*, that allows textual representations to benefit from visual properties without degrading the semantics brought by the textual objective $\mathcal{L}_\mathcal{T}$.

6.2.2 Grounding space and objectives

In this section, we introduce more formally the grounded space and the different information (cluster and perceptual) captured in the grounding loss $\mathcal{L}_\mathcal{G}$.

Grounded space The grounded space relaxes the assumption that textual and visual representations should be guided by one-to-one correspondences. It rather assumes that the structure of the textual space might be partially modeled on the structure of the visual space. Thus, instead of directly applying the grounding objectives on a sentence s embedding, we propose to train the grounding objective $\mathcal{L}_\mathcal{G}$ on an intermediate space called *grounded space*. Practically, we use a projection $g(s;\theta^i_g)$ of a sentence s from the textual space to the grounded space. We denote it $g(s)$ for simplicity, where g is a multi-layer perceptron with input $s = F^t(S;\theta^t)$ and parameters θ^i_g ($\theta^i_g \subset \theta^i$).

Cluster information (C_g) The cluster information leverages the fact that two sentences describe, or not, the same underlying reality. In other words, the goal is to measure if two sentences are *visually equivalent* (assumption (1) in Section 3.1) without considering the content of related images. For convenience, two sentences are said to be *visually equivalent* (resp. *visually different*) if they are associated

with the same image (resp. different images), i.e. if they describe the same (resp. different) underlying reality. We call *cluster* a set of visually equivalent sentences. For instance, in Figure 6.1, sentences *The tenniswoman starts on her serve* and *The woman plays tennis* are visually equivalent and belong to the same cluster.

Our hypothesis is that *the similarity between visually equivalent sentences* (s, s^+) *should be higher than visually different sentences* (s, s^-). We translate this hypothesis into the constraint in the grounded space: $\cos(g(s), g(s^+)) \leq \cos(g(s), g(s^-))$. Following (Karpathy et al. 2017; Carvalho et al. 2018), we use a max-margin ranking loss to ensure the gap between both terms is higher than a fixed margin γ (cf. red elements in Figure 6.1) resulting in the cluster loss $\mathcal{L}_\mathcal{C}$:

$$\mathcal{L}_C = \sum_{(s,s^+,s^-)} \left[\gamma - \cos(g(s), g(s^+)) + \cos(g(s), g(s^-))\right]_+ \tag{6.2}$$

where s^+ (resp. s^-) is a randomly sampled visually equivalent (resp. different) sentence to s. This loss function is also used in the cross-modal retrieval literature to enforce structure-preserving constraints between sentences describing a same image (L. Wang et al. 2016).

Perceptual information (P_g) The cluster hypothesis alone ignores the structure of the visual space and only uses the visual modality as a proxy to assess if two sentences are visually equivalent or different. Moreover, the ranking loss $\mathcal{L}_\mathcal{C}$ simply drives apart visually different sentences in the representation space, which can be a problem when two images have a closely related content. For instance, the baseball and tennis images in Figure 6.1 may be different, but they are both sports images, and thus their corresponding sentences should be somehow close in the grounded space. Finally, it supposes that we have a dataset of images associated with several captions.

To cope with these limitations, we consider the structure of the visual space and use the content of images. The intuition is that the structure of the textual space should be modeled on the structure of the visual one to extract visual semantics. We choose to preserve *similarities* between related elements across spaces (cf. green elements in Figure 6.1). We thus assume that *the similarity between two sentences in the grounded space should be correlated with the similarity between their corresponding images in the visual space.* We translate this hypothesis into the perceptual loss $\mathcal{L}_\mathcal{P}$:

$$\mathcal{L}_\mathcal{P} = -\rho(\{sim^{\text{text}}_{k_1,k_2}\}, \{sim^{\text{im}}_{k_1,k_2}\}) \tag{6.3}$$

where ρ is the Pearson correlation, $sim^{\text{text}}_{k_1,k_2} = \cos(g(s_{k_1}), g(s_{k_2}))$ and $sim^{\text{im}}_{k_1,k_2} = \cos(i_{k_1}, i_{k_2})$ are respectively textual and visual similarities computed over several randomly sampled pairs of matching sentences and images.

Grounded loss Taking altogether, the grounded space and cluster/perceptual information leads to the grounding objective $\mathcal{L}_{\mathcal{G}}(\theta^t, \theta^i)$ as a linear combination of the aforementioned objectives:

$$\mathcal{L}_{\mathcal{G}}(\theta^t, \theta^i) = \alpha_C \mathcal{L}_C(\theta^t, \theta^i) + \alpha_P \mathcal{L}_P(\theta^t, \theta^i) \tag{6.4}$$

where α_C and α_P are hyper-parameters weighting contributions of $\mathcal{L}_C$ and $\mathcal{L}_P$. θ^i corresponds to all the grounding-related parameters, i.e. those of the image encoder F^i and of the projection function g (i.e., θ^i_g).

6.3 Evaluation protocol

6.3.1 Datasets

Textual dataset. Following (Kiros et al. 2015; Felix Hill et al. 2016), we use the Toronto BookCorpus dataset as the textual corpus. This corpus consists of 11K books, and 74M ordered sentences, with an average of 13 words per sentence.

Visual dataset. We use the MS COCO (T. Lin et al. 2014a) dataset as the visual corpus. This image captioning dataset consists of 118K/5K/41K (train/val/test) images, each with five English descriptions. Note that the number of sentences in the training set of COCO (590K sentences) only represents 0.8% of the sentence data in BookCorpus, which is negligible, and the additional textual training data cannot account for performance discrepancies between textual and grounded models.

6.3.2 Baselines and Scenarios

In the experiments, we focus on one of the most established sentence models: SkipThought (noted **T**) as the textual baseline: the parameters of the sentence embedding model are obtained by minimizing $\mathcal{L}_{\mathcal{T}}$. Then, we derive several baselines and scenarios based on **T**, each representing a different approach of grounding. Since our focus is to study the impact of grounding on sentence representations, all baselines and scenarios share the same representation dimension $d_t = 2048$ and are trained on the same datasets (cf. Section 6.3.1). We also report a textual model of dimension $\frac{d_t}{2}$ that we call $\mathbf{T}_{1024}$, to compare with the GroundSent model of (Kiela et al. 2018).

Model Scenarios. We test variants of our grounding model presented in Section 6.2, all based on **T**: $\mathbf{T} + \mathbf{C}_g$, $\mathbf{T} + \mathbf{P}_g$, $\mathbf{T} + \mathbf{C}_g + \mathbf{P}_g$, where $\mathbf{C}_g$ (resp. $\mathbf{P}_g$) represents the loss $\mathcal{L}_C$ (resp. $\mathcal{L}_P$). We also consider scenarios where g equals the identity

function (no grounded space), which we note $\mathbf{C}_{id}$, $\mathbf{P}_{id}$, $\mathbf{C}_{id} + \mathbf{P}_{id}$, etc. Finally, we also performed preliminary analysis learning only from the visual modality: $\mathbf{C}_{g/id}$, $\mathbf{P}_{g/id}$, $\mathbf{C}_{g/id} + \mathbf{P}_{g/id}$.

Baselines. We adapt two classical multimodal word embedding models for sentences. Accordingly, models from the two existing model families are considered: *Cross-modal Projection* (**CM**): Inspired by Angeliki Lazaridou et al. 2015, this baseline learns to project sentences in the visual space using a max-margin loss:

$$\sum_{(s,i_s,i^-)} \left\lfloor \gamma' + \cos(f(s), i^-) - \cos(f(s), i_s) \right\rfloor_+$$

where f is a MLP, γ' is a fixed margin, i_s is the image corresponding to the sentence s, and i^- is a non-matching image. Similarly to our scenarios, the sentence encoder is initialized with $\mathbf{T}$.

Sequential (**SEQ**): Inspired by Collell et al. 2017, we learn a linear regression model (W, b) to predict the visual representation of an image, from the representation of a matching caption. The grounded word embedding is the concatenation of the original SkipThought vector $\mathbf{T}$ and its predicted ("*imagined*") representation $W\mathbf{T} + b$, which is projected using a PCA into dimension d_t.

In both cases, the parameters to be learned, in addition to the sentence encoder, are the cross-modal projections – and the sentence representation is obtained by averaging word vectors.

GroundSent Model We re-implement the GroundSent models of Kiela et al. 2018, obtaining comparable results. The authors propose two objectives to learn a grounded vector: (a) Cap2Img: the cross-modal projections of sentences are pushed towards their respective images via a max-margin ranking loss, and (b) Cap2Cap: a visually equivalent sentence is predicted via a LSTM sentence decoder. The Cap2Both objective is a combination of these two objectives. Once the grounded vectors are learned, they are concatenated with a textual vector (learned via a SkipThought objective) to form the GS-Img, GS-Cap and GS-Both vectors.

6.3.3 Evaluation tasks and metrics

In line with previous works (Kiros et al. 2015; Felix Hill et al. 2016), we consider several benchmarks to evaluate the quality of our grounded embeddings:

Semantic relatedness. We use two semantic similarity benchmarks: Semantic Textual Similarity (STS) (D. M. Cer et al. 2017) and Sentences Involving Compositional Knowledge (SICK) (Marelli et al. 2014a), which consist of pairs of sentences

that are associated with human-labeled similarity scores. STS is subdivided into three textual sources: *Captions* contain concrete sentences describing daily-life actions, whereas the others contain more abstract sentences: news headlines in *News* and posts from user forums in *Forum*. The Spearman correlations are measured between the cosine similarity of our learned sentence embeddings and human-labeled scores.

Classification benchmarks. All extrinsic evaluations are carried out using the SentEval pipeline (Conneau et al. 2018). The tasks are the following: Multi-Perspective Question Answering (MPQA) (Wiebe et al. 2005), Movie Review (MR) (Pang et al. 2005), Subjectivity/Objectivity (SUBJ) (Pang et al. 2004), Customer Reviews (CR) (M. Hu et al. 2004), binary sentiment analysis on Stanford Sentiment Treebank (SST) (Socher et al. 2013), paraphrase identification on Microsoft Research Paraphrase (MSRP) (Dolan et al. 2004) as well as two entailment classification benchmarks: Stanford Natural Language Inference (SNLI) (Bowman et al. 2015) and SICK (Marelli et al. 2014b). For each dataset, a logistic regression classifier is learned from the extracted sentence embeddings, and we report the classification accuracy.

Structural measures. To probe the learned grounded space, we define structural measures, and report their values on the validation set of MS COCO (5K images, 25K captions). First, we report the *mean Nearest Neighbor Overlap* (`mNNO`) metric, as defined in G. Collell et al. 2018, that indicates the proportion of shared nearest neighbors between image representations and their corresponding captions in their respective spaces. To study *perceptual information*, we define ρ_{vis}, the Pearson correlation $\rho(\cos(s, s'), \cos(v_s, v_{s'}))$ between images and their corresponding sentences' similarities. For *cluster information*, we introduce $C_{intra} = \mathbb{E}_{v_s = v_{s'}}[cos(s, s')]$, which measures the homogeneity of each cluster, and $C_{inter} = \mathbb{E}_{v_s \neq v_{s'}}[cos(s, s')]$, which measures how well clusters are separated from each other.

6.3.4 Implementation details

Images are processed using a pretrained Inception-v3 network (Szegedy et al. 2016b) ($d_i = 2048$). The model is trained with ADAM (Kingma et al. 2014a) and a learning rate $l_r = 8.10^{-4}$. As done in Kiros et al. 2015, our sentence encoder is a GRU with a vocabulary of 20K words, represented in dimension 620; we perform vocabulary expansion at inference. All hyperparameters are tuned using the Pearson correlation measure on the validation set of the SICK benchmark: $\gamma = \gamma' = 0.5$, $\alpha_C = \alpha_P = 0.01$, $d_g = 512$; functions f and g are 2-layer MLP. As done in (Kiela et al. 2018), we set $d_t = 2048$.

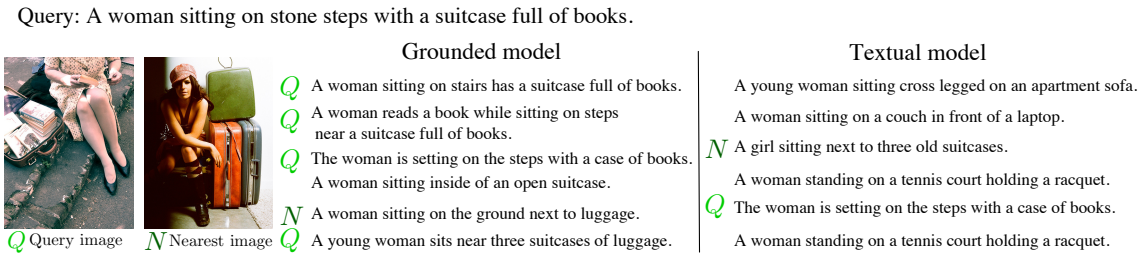

Figure 6.2 – Nearest neighbors of a selected sentence in the validation set of MS COCO, for both grounded and purely textual models. Q is the query image, N is the nearest neighbor of Q in the visual space. Sentences that are caption of Q or N are prefixed with Q or N.

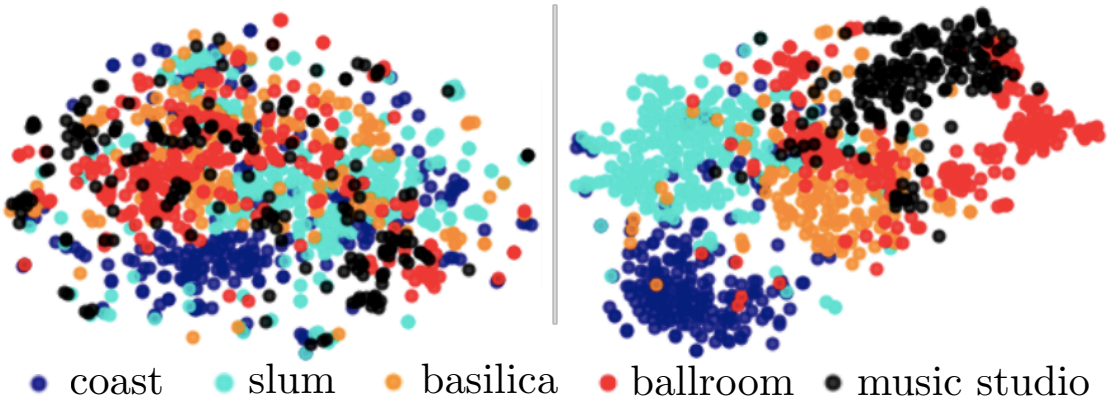

Figure 6.3 – t-SNE visualization on CMPlaces sentences for a set of randomly sampled visual scenes. Left: textual model $\mathbf{T}$. Right: grounded model $\mathbf{C}_g + \mathbf{P}_g$.

6.4 Experiments and Results

Our main objective is to study the contribution brought by the visual modality to the grounded sentence representations. Hence, we do not attempt to outperform purely textual sentence encoders from the literature.

We show that textual models can benefit from grounding approaches without requiring any changes to the original textual objectives $\mathcal{L}_T$. We report quantitative and qualitative insights (Section 6.4.1), and quantitative results on the SentEval benchmark (Section 6.4.2).

		Structural measures				Semantic relatedness				
Space	Model	mNNO	ρ_{vis}	C_{inter}	C_{intra}	STS/All	STS/Cap	STS/News	STS/Forum	SICK
Textual	**T**	10.0	4.1	54.2	70.1	30	41	36	21	51
	CM (text)	24.2	12.8	41.7	74.8	52	76	42	**37**	55
	$\mathbf{P}_{id}$	21.1	**37.9**	42.2	69.3	45	66	41	34	54
	$\mathbf{C}_{id}$	27.5	10.5	**2.9**	**84.7**	60	83	45	20	55
	$\mathbf{C}_{id} + \mathbf{P}_{id}$	**27.9**	25.8	6.7	82.6	**61**	**84**	**46**	28	**57**
Visual	**CM** (vis.)	27.1	19.2	1.5	85.8	56	78	40	34	55
Grounded	$\mathbf{P}_g$	21.3	**32.4**	43.9	73.3	45	66	41	**37**	53
	$\mathbf{C}_g$	28.6	9.4	**1.1**	**88.5**	62	83	46	29	59
	$\mathbf{C}_g + \mathbf{P}_g$	**28.9**	29.1	4.7	87.5	**63**	**84**	**48**	33	**60**

Table 6.1 – **Intrinsic evaluations** carried out on the grounded space for models with g = MLP; the textual space for **T**, **CM** (text) and models with $g = id$. The visual space for **CM** (vis). **CM** (text) and **CM** (vis) refer to the same model, the only difference is the space in which the measures are calculated (given in the parenthesis)

6.4.1 Study of the grounded space

We study the impact of the various grounding hypotheses on the structure of the grounded space, using intrinsic measures. In Table 6.1, we report the structural measures and the semantic relatedness scores of the baselines, namely **T** and **CM**, and on the various scenarios of our model. The textual loss is discarded to isolate the effect of the different grounding hypotheses.

The impact of grounding We investigate the effect of grounding on sentence representations. Results highlight that all grounded models improve over the baseline **T**. Moreover, our model $\mathbf{C}_g + \mathbf{P}_g$ is generally the most effective regarding the mNNO measure and semantic relatedness tasks.

Influence of concreteness To understand in which cases grounding is useful, we compute the average visual concreteness $\bar{c}$ of the STS benchmark, which is divided in three categories (*Captions*, *News*, *Forum*). This is done by using a concreteness dataset built by Brysbaert et al. 2013 consisting of human ratings of concreteness (between 0 and 5) for 40,000 English words; for a given benchmark, we compute the sum of these scores and average over all words that are in the concreteness dataset. The performance gain Δ between $\mathbf{C}_g + \mathbf{P}_g$ and **T** are observed when the visual concreteness $\bar{c}$ is high: for *Captions* ($\bar{c} = 3.10$), the improvement is substantial: ($\Delta = +43$); for benchmarks with a lower concreteness (*News* with $\bar{c} = 2.61$ and *Forum* with $\bar{c} = 2.39$), the improvement is smaller ($\Delta = +12$). Thus,

grounding brings useful complementary information, especially for concrete sentences.

t-distributed Stochastic Neighbor Embedding (t-SNE) visualization This finding is also supported by a qualitative experiment showing that grounding sentences clusters together similar visual situations. Using sentences from Cross-Modal Places (CMPlaces) (Castrejon et al. 2016), which describe visual scenes (e.g., *coast*, *shoe-shop*, *plaza*, etc.) and are classified in 205 scene categories, we randomly sample 5 visual scenes and plot in Figure 6.3 the corresponding sentences using t-SNE (Maaten et al. 2008). We notice that our grounded model is better able to cluster sentences that have a close visual meaning than the text-only model. This is reinforced by the structural measures computed on the five clusters of Figure 6.3: $C_{inter} = 19, C_{intra} = 22$ for $\mathbf{T}$, $C_{inter} = 11, C_{intra} = 27$ for $\mathbf{C}_g + \mathbf{P}_g$. Indeed, C_{inter} (resp. C_{intra}), is lower (resp. higher) for the grounded model $\mathbf{C}_g + \mathbf{P}_g$ compared to $\mathbf{T}$, which shows that clusters corresponding to different scenes are more clearly separated (resp. sentences corresponding to a given scene are more packed).

Nearest neighbors search Furthermore, we show in Table 6.2 that concrete knowledge acquired via our grounded model can also be transferred to abstract sentences. To do so, we manually build sentences using words with low concreteness (between 2.5 and 3.5) from the USF dataset (Nelson et al. 2004). Then, nearest neighbors are retrieved from the set of sentences of Flickr30K (Plummer et al. 2015). In this sample, we see that our grounded model is more accurate than the purely textual model to capture visual meaning. The observation that visual information propagates from concrete sentences to abstract ones is analogous to findings made in previous research on word embeddings (Felix Hill et al. 2014a).

Neighboring structure To illustrate the discrepancy on the `mNNO` metric observed between $\mathbf{C}_g + \mathbf{P}_g$ and $\mathbf{T}$, we select a query image Q in the validation set of MS COCO, along with its corresponding caption S; we display, in Figure 6.2, the nearest neighbor of Q in the visual space, noted N, and the nearest neighbors of S in the grounded space. With our grounded model, the neighborhood S is mostly made of sentences corresponding to Q or N.

Query	Textual model	Grounded model
Two people are in **love**	Two people are fencing indoors	A couple just got married and are taking a picture with family
A man is **horrified**	A man and a woman are smiling	A teenage boy wearing a cap looks irritated
This is a **tragedy**	A group of people are at a party	Men doing a war reenactment

Table 6.2 – **Qualitative study.** Nearest neighbor of a given query among Flickr30K sentences.

Hypotheses validation We now validate our hypotheses (cf. Section 6.2.1) on the grounded space, using the Cross-Modal Projection baseline (**CM**) and our model scenarios as outlined in Table 6.1. For fair comparison, metrics for the baseline **CM** are estimated either on the visual or the textual space depending on whether our models rely on the grounded space (g) or not (id). These results correspond to the rows **CM** (text) and **CM** (vis.) in Table 6.1.

Results highlight that:

1. Using a grounded space is beneficial; indeed, semantic relatedness and `mNNO` scores are higher in the lower half of Table 6.1, e.g., $\mathbf{C}_g > \mathbf{C}_{id}$, $\mathbf{P}_g > \mathbf{P}_{id}$ and $\mathbf{C}_g + \mathbf{P}_g > \mathbf{C}_{id} + \mathbf{P}_{id}$;
2. Solely using cluster information leads to the highest C_{intra} and lowest C_{inter}, which suggests that $\mathbf{C}_\bullet$ is the most efficient model at separating visually different sentences;
3. Using only perceptual information in $\mathbf{P}_\bullet$ logically leads to highly correlated textual and visual spaces (highest ρ_{vis}), but the local neighborhood structure is not well preserved (lowest C_{intra});
4. Our model $\mathbf{C}_\bullet + \mathbf{P}_\bullet$ is better than **CM** at capturing cluster information (higher C_{intra}, lower C_{inter}) and perceptual information (higher ρ_{vis}). This also translates in a higher `mNNO` measure for $\mathbf{C}_\bullet + \mathbf{P}_\bullet$, leading us to think that the conjunction of both perceptual and cluster information leads to high correlation of modalities, in terms of neighborhood structure. Moreover, this high `mNNO` score results in better performances for our model $\mathbf{C}_\bullet + \mathbf{P}_\bullet$ in terms of semantic relatedness.

6.4.2 Evaluation on transfer tasks

We now focus on extrinsic evaluation of the embeddings. Table 6.3 reports evaluations of our baselines and scenarios on SentEval (Conneau et al. 2018), a classical benchmark used for evaluating sentence embeddings. Before further analysis, we find that our grounded models systematically outperform the textual baseline **T**, on all benchmarks, which shows the first substantial improvement brought by grounding and visual information in a sentence representation model. Indeed, models GS-Cap, GS-Img and GS-Both from (Kiela et al. 2018), despite improving over $\mathbf{T}_{1024}$, perform worse than the textual model of the same dimension **T** — this is consistent with what they report in their paper.

Our results interpretation is the following:

1. our joint approach shows superior performances over the sequential one, confirming results reported at the word level (Eloi Zablocki et al. 2018b). Indeed, both sequential models, GS models (Kiela et al. 2018) and **SEQ** (inspired from (Collell et al. 2017)) are systematically worse than our grounded models for all benchmarks.

Model		MR	CR	SUBJ	MPQA	MRPC	SST	SNLI	SICK	AVG
Kiros et al. 2015†	$\mathbf{T}_{1024}$	72.7	75.2	90.6	84.7	71.8/79.2	76.2	68.8	79.3	77.4
Kiela et al. 2018†	GS-Cap	72.0	76.8	90.7	85.5	72.9/80.6	76.7	73.7	**82.9**	78.4
Kiela et al. 2018†	GS-Img	74.5	79.3	90.8	87.8	73.0/80.3	80.0	72.2	80.9	79.8
Kiela et al. 2018†	GS-Both	72.5	75.7	90.7	85.4	72.9/81.3	76.7	72.2	81.4	78.4
Kiros et al. 2015†	$\mathbf{T}$	75.9	79.2	92.0	86.7	72.2/80.2	81.8	72.0	81.1	80.1
Lazaridou et al. 2015‡	$\mathbf{T} + \mathbf{CM}$	77.6	81.4	92.6	88.3	73.5/81.1	82.0	73.0	81.4	81.1
Collell et al. 2017‡	**SEQ**	76.1	79.8	92.5	86.7	70.0/79.5	81.7	67.3	76.7	78.9
Model scenarios	$\mathbf{T} + \mathbf{P}_{id}$	77.5	81.5	92.7	88.4	**73.7**/81.3	82.4	72.4	81.1	81.2
	$\mathbf{T} + \mathbf{P}_{g}$	**77.8**	**81.8**	**93.0**	88.1	73.3/**81.6**	**83.5**	72.8	82.2	**81.6**
	$\mathbf{T} + \mathbf{C}_{id}$	77.5	81.6	92.8	88.3	72.9/80.5	82.2	73.1	82.3	81.3
	$\mathbf{T} + \mathbf{C}_{g}$	77.3	81.5	92.8	**88.6**	73.6/81.1	82.6	**74.1**	82.6	**81.6**
	$\mathbf{T} + \mathbf{C}_{id} + \mathbf{P}_{id}$	77.3	81.2	**93.0**	88.4	73.0/80.6	82.5	73.5	82.1	81.4
	$\mathbf{T} + \mathbf{C}_{g} + \mathbf{P}_{g}$	77.4	81.5	**93.0**	88.1	73.2/80.9	82.7	73.9	**82.9**	**81.6**

Table 6.3 – **Extrinsic evaluations with SentEval** All models give sentences in dimension $d_t = 2048$ (except $\mathbf{T}_{1024}$). 'AVG' stands for the average accuracies reported in the other columns. Models noted '†' have been re-implemented (we report higher scores than the one given in the original papers). Models noted '‡' are baselines which have been adapted to the case of sentences.

2. Preserving the structure of the visual space is more effective than learning cross-modal projections; indeed, all our models outperform $\mathbf{T} + \mathbf{CM}$ on average ('AVG' column).

3. Making use of a grounded space yields slightly improved sentence representations. Indeed, our models that use the grounded space (g = MLP) can take advantage of more expression power provided by the trainable g than models which integrate grounded information directly in the textual space ($g = id$).

4. Among our model scenarios, $\mathbf{T} + \mathbf{P}_g$ has maximal scores on the most tasks; however, it shows lower scores on SNLI and SICK, which are entailment tasks. Models using cluster information $\mathbf{C}_g$ are naturally more suited for these tasks and hence obtain higher results.

5. The combined model $\mathbf{T} + \mathbf{C}_g + \mathbf{P}_g$ shows a good balance between classification and entailment tasks.

6.5 Conclusion

6.5.1 Summary of the contributions

We proposed a multimodal model aiming at preserving the structure of visual and textual spaces to learn grounded sentence representations. Our contributions include:

1. leveraging both perceptual and cluster information
2. using an intermediate grounded space enabling to relax the constraints on the textual space.

Our approach is the first to report consistent positive results against purely textual baselines on a variety of natural language tasks.

We can answer to the Chapter Question:

- **CQ**: Visual semantics can be transferred within sentence representations, as we showed both quantitatively and qualitatively.

6.5.2 Perspectives

In this section, we present research perspectives following the contribution made in this Chapter.

Using videos to ground language As human gain a grounded understanding of the meaning of words and sentences, they always witness *actions* rather than still images. Moreover, the understanding of action verbs is done with the temporal dimension, in addition to the visual modality. Thus, we can hint that *videos*, rather than *images*, may lead to better grounded sentence representations. Inspiration from recent developments of Multimodal Neural Language Models, such as VideoBERT (C. Sun et al. 2019), may lead to better grounded sentence representations, able to better capture temporal aspects of language.

CHAPTER 7

CONCLUSION

Contents

7.1 Summary and Contributions

In the present thesis, we addressed central themes of Multimodal Machine Learning, that we organized in the Introduction in three global Research Questions:

- **RQ1**: Can vision help to refine language understanding ?
- **RQ2**: Can language help to refine visual understanding ?
- **RQ3**: Can modalities be translated into one another?

In each Chapter, we covered a specific aspect of these RQ; the Chapter Questions addressed throughout this thesis are given in Table 7.1.

In Chapter 3, we showed that textual representations, like the Word2Vec model (Mikolov et al. 2013b), contain visual information about objects, that can be exploited in a Zero-Shot Learning (ZSL) setting. We showed, by proposing a new task — context-aware ZSL — along with a corresponding model, that textual representations contain information on (i) the visual appearance of objects, (ii) the visual context around objects in images and (iii) the frequency of objects in images, thus answering the question: *What visual information is encoded in word embeddings?*. In context-aware ZSL, the goal is to determine the class of an object, delimited by a bounding box in the image, by taking into account the other objects of the image. To understand the influence of (i), (ii) and (iii), we formulate the ZSL problem using Bayesian modeling, with a visual, contextual and prior component that have separate goals and distinct learning objectives. We thus answer: *How*

Chapter	RQ1	RQ2	RQ3
Leveraging Visual Knowledge within Language for Computer Vision		What visual information is encoded in word embeddings ?	How can cross-modal models be more interpretable ?
Leveraging Weak/Non-existent Cross-Modal Supervision	Can multimodal tasks benefit from multimodal data when cross-modal supervision is weak, or non-existent ?		
On the Cross-Modal Transferability of Language Models	Is vision helpful for Question Generation?	Do language models contain visual information ?	How do modalities affect Question Generation?
Grounding language in vision: the case of Sentences	Can visual semantics be trans--ferred to sentence representations ?		

Table 7.1 – Research Questions addressed in the present thesis

can cross-modal models be more interpretable? by proposing a modular model which enables an accurate understanding of the model's prediction. We show that using contextual information leads to a 22% relative improvement on the Mean First Relevant (MFR) metric, compared to the standard DeViSE model (Frome et al. 2013), that only focuses on the visual appearance of objects. This work and results have been published at ICML 2019.

In Chapter 4, we answer the question: *Can multimodal tasks benefit from multimodal data when cross-modal supervision is weak, or non-existent?* positively by tackling various multimodal tasks in settings with weak supervision: (i) Transductive Zero-Shot Learning (T-ZSL) on the full ImageNet dataset (20K unseen classes), (ii) Zero-Shot Cross-Modal Retrieval on MS COCO and (iii) Visual Grounding of Language. To do so, we adapt CycleGAN (J. Zhu et al. 2017) to align textual and visual distributions when no supervision is available. Our model for T-ZSL, that we call Cross-Modal CycleGAN, combines a CycleGAN objective, trained on unseen data, with a supervised objective, trained on seen data. We show that using adversarial learning enables to exploit unsupervised multimodal data, and compares favorably to other approaches when the number of classes is high. Cross-Modal CycleGAN (i) obtains state-of-the-art results on the challenging ImageNet dataset, (ii) learns some cross-modal alignment on MS COCO when no correspondence is known between images and sentences, and (iii) learns meaningful grounded textual representations without supervision between words and images. This work is currently under review at IJCAI 2020.

In Chapter 5, we explore the cross-modal transferability capabilities of Language Models. Following recent works that show that BERT (Devlin et al. 2019) contains abstractions that generalize across languages (Artetxe et al. 2019), our goal is to show that it also contains abstractions that generalize across modalities. To do so, we tackle a multimodal task: Visual Question Generation (VQG) and design a multimodal version of BERT, able to leverage both textual and visual inputs. We answer: *Do language models contain visual information?* positively. Indeed, in our model, we integrate visual elements in BERT by treating vision at the same level

as language: we use a simple linear layer to project visual representations in BERT embedding layer. We show that this model is sufficient to obtain substantial results on VQG. We also answer: *Is vision helpful for Question Generation?* positively, with quantitative and qualitative results: we showed that the visual modality refines question predictions and that the attention process of BERT can focus on relevant elements of the image. Finally, we demonstrate that using both modalities as input (an image along with its corresponding caption) leads to the highest-quality questions, which answers: *How do modalities affect Question Generation?* Our model obtains state-of-the-art performances on two standard VQG datasets. This work is currently under review at ICML 2020.

In Chapter 6, we tackle Visual Grounding of Language (VGL) at the *sentence* level. By listing the differences between sentences and words: — sentences have a wide variety of synonym sentences carrying th same visual information, can be visually ambiguous and can carry non-visual information — we derive two training objectives to learn grounded sentence representations: a (i) cluster objective, that bring together sentences that contain the equivalent visual information, and a (ii) perceptual objective, that uses the content of images to preserve the structure of the visual space within the grounded representation space. In addition, we propose to use an intermediate grounded space to relax the implicit assumption that modalities should have a one-to-one correspondence. Our approach is the first to report consistent improvement over purely textual baselines on a variety of textual tasks (+1.3 on average on SentEval), thus answering positively: *Can visual semantics be transferred to sentence representations?* This work has been published at EMNLP 2019.

7.2 Open questions and perspectives

7.2.1 Extensions and perspectives of our approaches

Designing meaningful metrics to compare text and vision When learning a *shared* multimodal space (Section 2.1.3.2), measuring the quality of the cross-modal alignment is straightforward, using the tools of Information Retrieval e.g, Recall/Precision/F1 metrics, or First Relevant (FR)/MFR.

However, very few work have explored ways to measure the differences between a visual and a textual representation spaces, when they are *separated*. As an example, G. Collell et al. 2018 proposed to $mNNO$ metric: $mNNO$ measures the proportion of nearest neighbors shared between two corresponding from both modalities, in average. In Chapter 6 and Chapter 4, we use the ρ_{vis} metric (Equation 4.3), which measures the correlation between the similarities of a pair of corresponding elements across modalities.

Such metrics could lead to interesting applications. First, new methods could be invented to learn cross-modal alignments without explicit projection between spaces — Chapter 6 is a first attempt toward that goal. At longer term, we could have a quantitative understanding of the semantic differences between text and vision using carefully-designed metrics.

Learning grounded relation representations Learning representations for relations — in the form *(subject,predicate,object)* (s, p, o) — has emerged as an important issue in the last decade, with the TransE (A. Bordes et al. 2013), TransH (Z. Wang et al. 2014) and TransR (Y. Lin et al. 2015) models, which learn relation embeddings for Knowledge Bases. Focusing on *visual* relations — either spatial elements like *below*, *under*, *behind* or action verbs such as *eating*, *hugging* — would bring important improvements to many multimodal task that require visual reasoning, as in Visual Question Answering (VQA), VQG, or Visual Relationship Detection (VRD). For example, one could imagine a framework in which (i) a Faster R-CNN (S. Ren et al. 2017) detects objects in images, (ii) a relation module detects most probable relations between objects, and (iii) this information is fed to a decoder that generates a relevant caption, a question or an answer.

Learning grounded relation representations could be done using Visual Genome data (Krishna et al. 2017), as this dataset contains fine-grained annotations of images: labeled bounding boxes along with relations between objects. Once meaningful representations for objects, predicates and subjects have been learned, probabilities of triplets (s, p, o) could be computed. Thus, it would be possible to estimate the plausibility of certain relations compared to others, following works such as F. Sadeghi et al. 2015. For example, we could estimate that *(man,riding,horse)* is more probable than *(dog,riding,horse)*, *(man,eating,horse)* or *(man,riding,dog)*.

7.2.2 Research perspectives

Leveraging abstract scenes One key-insight in recent works is that low-level information in not needed in order to learn common sense, but rather high level semantic features. Hence, working with abstract scenes (clipart images where the position, pose and attributes of objects is known) is an interesting way to gain accurate high-level information (Kottur et al. 2016; Vedantam et al. 2015b), that may be difficult to acquire from noisy natural images. To do so, a toy dataset is created with an ontology of objects and possible actions.

Various applications have been proposed: learning occurence/co-occurence of objects (Zitnick et al. 2013; Zitnick et al. 2016), dynamics of objects (Fouhey et al. 2014), fine-grained interactions between pairs of people (Antol et al. 2014), classify common sense assertions as plausible or not (Vedantam et al. 2015b; Kottur et al. 2016) and imagine abstract scenes corresponding to text (X. Lin et al.

2015; Kottur et al. 2016). The latter task is especially interesting to incorporate visual common-sense in NLP models: it has so far been tackled using simple tools (considering sentences as a sum of word embeddings), and would benefit from latest improvements in Natural Language Processing (NLP), such as the BERT model (Devlin et al. 2019).

Fine-grained understanding of the visual content of textual representations There is still work to be done to design fine-grained tasks and benchmarks to understand in more detail what type of visual information is contained in textual representations. An interesting tentative is the feature-norm task using the McRae dataset (see Section 2.1.1.1), where an ontology of properties is defined, that are grouped in domains such as *Tactile*, *Color* or *Shape*. With this task, a high score in each regarding an aspect, for example *Color*, hints that textual representations encode useful information about the color of objects.

A step further would be to understand whether word representations encode relationships between objects. For example, we can propose the following task: given two objects s and p, predicting the most probable relationship between these objects. This task would require to train a model: for example, a 1-layer MLP classifier, that takes as input the concatenation of the representations t_s and t_p of s and p, and outputs a distribution over a finite set of pre-defined relationships (e.g., *riding*, *below*, *eating*).

At the sentence level, the only task commonly used to assess the visual content of sentence representations is the Cross-Modal Retrieval (CMR) task, as done in Kiros et al. 2015. The construction of a dataset containing sentences labeled with visual categories, like the McRae dataset for words, would be helpful to refine our understanding of grounded sentence representations, and would extend Chapter 6.

At another granularity level, by taking inspiration from our VQG method presented in Chapter 5, Neural Language Model (NLM)s can be evaluated by (i) integrating visual content with a cross-modal linear layer (that projects visual tokens into the embedding layer), (ii) leaving the weights of the NLM constant, and (iii) learning the cross-modal projection on a multimodal task. In Chapter 5, we considered VQG; other tasks can be used, like Image Captioning, or VQA. This would enable to evalute the capacity of NLMs to tackle specific visual tasks, and thus gain a deeper understanding of the visual content of large Language Models.

7.2.3 Longer-term research directions

Image Synthesis using human cues As explained in Section 2.4.2, Text-to-Image synthesis is a difficult task that is often restricted to very specific domains (e.g., birds or flowers pictures). While there is still little work on general-domain Text-to-Image synthesis, like MS COCO data, B. Li et al. 2019 proposed an interesting

direction: a word-level discriminator is used to provide fine-grained supervisory feedback to the image generator, in order to control parts of the image synthesis.

A step further would be to provide a human feedback, in natural language, to orient image generation. Thus, using human cues, the model would be oriented toward the image that the human subject has in mind. The ideal process is the following: from an input sentence s, an image $I_s = G(s)$ is generated; then, given I_s, a human feedback sentence s' is given to the model, from which the model produces $I_{s'} = G(s'|I_s, s)$, by refining the image I_s using feedback s', and so on. To learn such a model, a dataset containing $s, I_s, s', I_{s'}$ examples would have to be built, by first proposing a pair of sentences s, s' and then taking two pictures of the same scene, with some variations between I_s and $I_{s'}$.

Generating visually plausible sequences Many NLP tasks, such as Question Answering or Dialog Systems, involve reasoning and common-sense knowledge. As (i) the model cannot use a visual support, contrarily to their multimodal counterparts (Visual Question Answering and Visual Dialog), and (ii) training is performed only on textual data, the model may lack visual common-sense and generate un-relevant questions/answers in some situations.

Generally, when the decoder of a NLP model sequentially generates a sentence, a beam search (Freitag et al. 2017; Wiseman et al. 2016) is performed, where the principle is to keep a limited set of node explorations. At the end, the criterion to select the most probable sequence is a purely textual criterion: it has to maximize a probability $P(x_1, ..., x_n)$ given by a neural language model (generally trained on textual data). At this stage, adding an additional *visual criterion* might lead the model to generate more visually-plausible sentences. This could be done via an evaluation of the plausibility of the visual relations $P(s, p, o)$ for triplets s, p, o present in generated sentences (see Section 7.2.1). For example, in a conversational setting, the question *Did you enjoy riding horses when you were a kid?* may get a stronger visual probability than the question *Did you enjoy riding dogs when you were a kid?*.

Moreover, visual logic derived from abstract scenes could be used. Indeed, we proposed, earlier in this Section, a perspective where simple abstract scenes could be built from sentences with a concrete/visual content. Being able to navigate between sentences and abstract scenes would enable NLP models to gain visual common-sense: by (i) translating a sentence into an abstract scene, (ii) evaluate the plausibility of the scene, and what are its probable outcomes and (iii) translate back into the space of sentences.

BIBLIOGRAPHY

Ahn, Luis von and Laura Dabbish (2005). "ESP: Labeling Images with a Computer Game". In: *Knowledge Collection from Volunteer Contributors, Papers from the 2005 AAAI Spring Symposium, Technical Report SS-05-03, Stanford, California, USA, March 21-23, 2005*, pp. 91–98 (cit. on pp. 3, 36).

Akaho, Shotaro (2006). "A kernel method for canonical correlation analysis". In: *CoRR* abs/cs/0609071. arXiv: cs/0609071 (cit. on p. 53).

Akata, Z., F. Perronnin, Z. Harchaoui, and C. Schmid (2016). "Label-Embedding for Image Classification". In: *TPAMI* (cit. on p. 58).

Akata, Z., S. E. Reed, D. Walter, H. Lee, and B. Schiele (2015). "Evaluation of output embeddings for fine-grained image classification". In: *CVPR* (cit. on p. 47).

Alberti, Chris, Jeffrey Ling, Michael Collins, and David Reitter (2019). "Fusion of Detected Objects in Text for Visual Question Answering". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pp. 2131–2140 (cit. on p. 32).

Almahairi, A., S. Rajeshwar, A. Sordoni, P. Bachman, and A. C. Courville (2019). "Augmented CycleGAN: Learning Many-to-Many Mappings from Unpaired Data". In: *ICML 2018* (cit. on pp. 52, 83).

Aly, Ahmed, Kushal Lakhotia, Shicong Zhao, Mrinal Mohit, Barlas Oguz, Abhinav Arora, Sonal Gupta, Christopher Dewan, Stef Nelson-Lindall, and Rushin Shah (2018). "Pytext: A seamless path from nlp research to production". In: *arXiv preprint arXiv:1812.08729* (cit. on p. 109).

Anderson, Peter, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang (2018). "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering". In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 6077–6086 (cit. on p. 104).

Andrew, Galen, Raman Arora, Jeff A. Bilmes, and Karen Livescu (2013). "Deep Canonical Correlation Analysis". In: *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pp. 1247–1255 (cit. on pp. 31, 53, 54).

Antol, Stanislaw, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh (2015a). "VQA: Visual Question Answering". In: *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pp. 2425–2433 (cit. on pp. 29, 42).

Antol, Stanislaw, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh (2015b). "VQA: Visual Question Answering". In: *CoRR* abs/1505.00468 (cit. on pp. 7, 12).

Antol, Stanislaw, C. Lawrence Zitnick, and Devi Parikh (2014). "Zero-Shot Learning via Visual Abstraction". In: *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV*, pp. 401–416 (cit. on p. 136).

Aries, Abdelkrime, Djamel Eddine Zegour, and Walid-Khaled Hidouci (2019). "Automatic text summarization: What has been done and what has to be done". In: *CoRR* abs/1904.00688. arXiv: 1904.00688 (cit. on p. 13).

Arroyo-Fernández, Ignacio, Carlos-Francisco Méndez-Cruz, Gerardo Sierra, Juan-Manuel Torres-Moreno, and Grigori Sidorov (2019). "Unsupervised sentence representations as word information series: Revisiting TF-IDF". In: *Computer Speech & Language* 56, pp. 107–129 (cit. on p. 20).

Artetxe, Mikel, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho (2018). "Unsupervised Neural Machine Translation". In: *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings* (cit. on p. 13).

Artetxe, Mikel, Sebastian Ruder, and Dani Yogatama (2019). "On the Cross-lingual Transferability of Monolingual Representations". In: *CoRR* abs/1910.11856. arXiv: 1910.11856 (cit. on pp. 102, 134).

Ba, Lei Jimmy, Kevin Swersky, Sanja Fidler, and Ruslan Salakhutdinov (2015). "Predicting Deep Zero-Shot Convolutional Neural Networks Using Textual Descriptions". In: *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pp. 4247–4255 (cit. on p. 48).

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2014). "Neural Machine Translation by Jointly Learning to Align and Translate". In: *CoRR* abs/1409.0473. arXiv: 1409.0473 (cit. on pp. 1, 13, 19, 118).

Baltrusaitis, Tadas, Chaitanya Ahuja, and Louis-Philippe Morency (2019). "Multimodal Machine Learning: A Survey and Taxonomy". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 41.2, pp. 423–443 (cit. on p. 101).

Banerjee, Satanjeev and Alon Lavie (2005). "METEOR: An automatic metric for MT evaluation with improved correlation with human judgments". In: *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72 (cit. on pp. 51, 108).

Bansal, A., K. Sikka, G. Sharma, R. Chellappa, and A. Divakaran (2018). "Zero-Shot Object Detection". In: *ECCV* (cit. on p. 62).

Baroni, Marco (2016). "Grounding Distributional Semantics in the Visual World". In: *Language and Linguistics Compass* 10.1, pp. 3–13 (cit. on pp. 5, 101, 118).

Barrault, Loıc, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank (2018). "Findings of the Third Shared Task on Multimodal Machine Translation". In: *Proceedings of the Third Conference on Machine Translation:*

Shared Task Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018, pp. 304–323 (cit. on p. 44).

Bell, S., C. L. Zitnick, K. Bala, and R. B. Girshick (2016). "Inside-Outside Net: Detecting Objects in Context with Skip Pooling and Recurrent Neural Networks". In: *CVPR* (cit. on p. 59).

Ben-younes, Hedi, Rémi Cadène, Matthieu Cord, and Nicolas Thome (2017). "MUTAN: Multimodal Tucker Fusion for Visual Question Answering". In: *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 2631–2639 (cit. on p. 30).

Bengio, S., J. Dean, D. Erhan, E. Ie, Q. V. Le, A. Rabinovich, J. Shlens, and Y. Singer (2013). "Using Web Co-occurrence Statistics for Improving Image Categorization". In: *CoRR* abs/1312.5697. arXiv: 1312.5697 (cit. on p. 68).

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Janvin (2003). "A Neural Probabilistic Language Model". In: *J. Mach. Learn. Res.* 3, pp. 1137–1155 (cit. on pp. 2, 22, 23).

Bengio, Yoshua, Patrice Y. Simard, and Paolo Frasconi (1994). "Learning long-term dependencies with gradient descent is difficult". In: *IEEE Trans. Neural Networks* 5.2, pp. 157–166 (cit. on p. 18).

Bernardi, Raffaella, Ruket Çakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank (2016). "Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures". In: *J. Artif. Intell. Res.* 55, pp. 409–442 (cit. on p. 12).

Bordes, Antoine, Nicolas Usunier, Alberto Garcıa-Durán, Jason Weston, and Oksana Yakhnenko (2013). "Translating Embeddings for Modeling Multi-relational Data". In: *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States*, pp. 2787–2795 (cit. on p. 136).

Bordes, Patrick, Eloi Zablocki, Laure Soulier, Benjamin Piwowarski, and patrick Gallinari (Nov. 2019). "Incorporating Visual Semantics into Sentence Representations within a Grounded Space". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 696–707 (cit. on p. 101).

Bowman, Samuel R., Gabor Angeli, Christopher Potts, and Christopher D. Manning (2015). "A large annotated corpus for learning natural language inference". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pp. 632–642 (cit. on pp. 21, 125).

Bruna, J., W. Zaremba, A. Szlam, and Y. LeCun (2014). "Spectral Networks and Locally Connected Networks on Graphs". In: *ICLR 2014* (cit. on p. 48).

Bruni, Elia, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran (2012a). "Distributional Semantics in Technicolor". In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*. ACL '12. Jeju Island, Korea: Association for Computational Linguistics, pp. 136–145 (cit. on p. 36).

Bruni, Elia, Nam Khanh Tran, and Marco Baroni (Jan. 2014). "Multimodal Distributional Semantics". In: *J. Artif. Int. Res.* 49.1, pp. 1–47 (cit. on pp. 16, 37, 95, 119).

Bruni, Elia, Jasper Uijlings, Marco Baroni, and Nicu Sebe (2012b). "Distributional Semantics with Eyes: Using Image Analysis to Improve Computational Representations of Word Meaning". In: *Proceedings of the 20th ACM International Conference on Multimedia*. MM '12. Nara, Japan: ACM, pp. 1219–1228 (cit. on p. 37).

Brysbaert, Marc, Amy Beth Warriner, and Victor Kuperman (Oct. 2013). "Concreteness ratings for 40 thousand generally known English word lemmas". In: *Behavior research methods* 46 (cit. on p. 127).

Bucher, M., S. Herbin, and F. Jurie (2016). "Improving Semantic Embedding Consistency by Metric Learning for Zero-Shot Classiffication". In: *ECCV* (cit. on p. 47).

Bulat, Luana, Douwe Kiela, and Stephen Clark (2016). "Vision and Feature Norms: Improving automatic feature norm learning through cross-modal maps". In: *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pp. 579–588 (cit. on p. 37).

Burgess, Curt and Kevin Lund (Mar. 1997). "Modelling Parsing Constraints with High-dimensional Context Space". In: 12 (cit. on p. 34).

Cadène, Rémi, Hedi Ben-younes, Matthieu Cord, and Nicolas Thome (2019). "MUREL: Multimodal Relational Reasoning for Visual Question Answering". In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 1989–1998 (cit. on p. 43).

Caglayan, Ozan, Loıc Barrault, and Fethi Bougares (2016). "Multimodal Attention for Neural Machine Translation". In: *CoRR* abs/1609.03976. arXiv: 1609.03976 (cit. on p. 44).

Caglayan, Ozan, Pranava Madhyastha, Lucia Specia, and Loıc Barrault (2019). "Probing the Need for Visual Context in Multimodal Machine Translation". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pp. 4159–4170 (cit. on p. 45).

Calixto, Iacer, Desmond Elliott, and Stella Frank (2016). "DCU-UvA Multimodal MT System Report". In: *Proceedings of the First Conference on Machine Translation,*

WMT 2016, colocated with ACL 2016, August 11-12, Berlin, Germany, pp. 634–638 (cit. on p. 44).

Calixto, Iacer and Qun Liu (2017). "Incorporating Global Visual Features into Attention-based Neural Machine Translation". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pp. 992–1003 (cit. on p. 44).

Carvalho, Micael, Rémi Cadène, David Picard, Laure Soulier, Nicolas Thome, and Matthieu Cord (2018). "Cross-Modal Retrieval in the Cooking Context: Learning Semantic Text-Image Embeddings". In: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, pp. 35–44 (cit. on pp. 32, 122).

Castrejon, Lluis, Yusuf Aytar, Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba (2016). "Learning Aligned Cross-Modal Representations from Weakly Aligned Data". In: *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*. IEEE (cit. on p. 128).

Cer, Daniel M., Mona T. Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia (2017). "SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation". In: *Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval@ACL 2017, Vancouver, Canada, August 3-4, 2017*, pp. 1–14 (cit. on pp. 20, 124).

Cer, Daniel, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil (2018). "Universal Sentence Encoder for English". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018*, pp. 169–174 (cit. on p. 20).

Changpinyo, S., W.-L. Chao, B. Gong, and F. Sha (2016). "Synthesized Classifiers for Zero-Shot Learning". In: *CVPR* (cit. on pp. 47, 48, 89, 90).

Changpinyo, S., W.-L. Chao, and F. Sha (2017). "Predicting Visual Exemplars of Unseen Classes for Zero-Shot Learning". In: *ICCV* (cit. on pp. 48, 90).

Chattopadhyay, Prithvijit, Deshraj Yadav, Viraj Prabhu, Arjun Chandrasekaran, Abhishek Das, Stefan Lee, Dhruv Batra, and Devi Parikh (2017). "Evaluating Visual Conversational Agents via Cooperative Human-AI Games". In: *Proceedings of the Fifth AAAI Conference on Human Computation and Crowdsourcing, HCOMP 2017, 23-26 October 2017, Québec City, Québec, Canada*, pp. 2–10 (cit. on p. 44).

Chen, Danqi, Adam Fisch, Jason Weston, and Antoine Bordes (2017). "Reading Wikipedia to Answer Open-Domain Questions". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pp. 1870–1879 (cit. on p. 13).

Chen, Xinlei and C. Lawrence Zitnick (2015). "Mind's eye: A recurrent visual representation for image caption generation". In: *IEEE Conference on Computer*

Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015, pp. 2422–2431 (cit. on pp. 40, 50).

Cheng, Zezhou, Qingxiong Yang, and Bin Sheng (2015). "Deep Colorization". In: *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pp. 415–423 (cit. on p. 25).

Cho, KyungHyun, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio (2014). "On the Properties of Neural Machine Translation: Encoder-Decoder Approaches". In: *CoRR* abs/1409.1259. arXiv: 1409.1259 (cit. on p. 19).

Chomsky, Noam (1980). "Rules and representations". In: *Behavioral and brain sciences* 3.1, pp. 1–15 (cit. on p. 34).

Chrupala, Grzegorz, Ákos Kádár, and Afra Alishahi (2015). "Learning language through pictures". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2: Short Papers*, pp. 112–118 (cit. on p. 39).

Chung, Y.-A., W.-H. Weng, S. Tong, and J. R. Glass (2018). "Unsupervised Cross-Modal Alignment of Speech and Text Embedding Spaces". In: *NeurIPS* (cit. on p. 84).

Clark, Kevin, Urvashi Khandelwal, Omer Levy, and Christopher D Manning (2019). "What Does BERT Look At? An Analysis of BERT's Attention". In: *arXiv preprint arXiv:1906.04341* (cit. on p. 112).

Collell, Guillem and M.-F. Moens (2018). "Do Neural Network Cross-Modal Mappings Really Bridge Modalities?" In: *ACL* (cit. on pp. 125, 135).

Collell, Guillem, Ted Zhang, and Marie-Francine Moens (2017). "Imagined Visual Representations as Multimodal Embeddings". In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pp. 4378–4384 (cit. on p. 101).

Collell, Teddy Zhang, and Marie-Francine Moens (2017). "Imagined visual representations as multimodal embeddings". In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*. AAAI (cit. on pp. 37, 38, 77, 94, 95, 119, 124, 129, 130).

Conneau, Alexis and Douwe Kiela (2018). "SentEval: An Evaluation Toolkit for Universal Sentence Representations". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. (Cit. on pp. 125, 129).

Dalal, Navneet and Bill Triggs (2005). "Histograms of Oriented Gradients for Human Detection". In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), 20-26 June 2005, San Diego, CA, USA*, pp. 886–893 (cit. on pp. 2, 26).

Das, Abhishek, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra (2017a). "Human Attention in Visual Question Answering: Do Humans and

Deep Networks Look at the Same Regions?" In: *Computer Vision and Image Understanding* 163, pp. 90–100 (cit. on pp. 40, 43).

Das, Abhishek, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, Stefan Lee, José M. F. Moura, Devi Parikh, and Dhruv Batra (2019). "Visual Dialog". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 41.5, pp. 1242–1256 (cit. on pp. 7, 12, 43).

Das, Abhishek, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra (2016). "Visual Dialog". In: *CoRR* abs/1611.08669. arXiv: 1611.08669 (cit. on pp. 43, 44).

Das, Abhishek, Satwik Kottur, José M. F. Moura, Stefan Lee, and Dhruv Batra (2017b). "Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning". In: *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 2970–2979 (cit. on p. 44).

Dash, Ayushman, John Cristian Borges Gamboa, Sheraz Ahmed, Marcus Liwicki, and Muhammad Zeshan Afzal (2017). "TAC-GAN - Text Conditioned Auxiliary Classifier Generative Adversarial Network". In: *CoRR* abs/1703.06412. arXiv: 1703.06412 (cit. on p. 51).

De Vega, M., A. Glenberg, and A. Graesser (Mar. 2012). *Symbols and embodiment: Debates on meaning and cognition*. English (US). Oxford University Press (cit. on p. 35).

Defferrard, M., X. Bresson, and P. Vandergheynst (2016). "Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering". In: *NeurIPS 2016* (cit. on p. 48).

Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li (2009). "ImageNet: A large-scale hierarchical image database". In: *CVPR* (cit. on pp. 2, 5, 25, 27, 45, 58).

Devereux, Barry, Lorraine Tyler, Jeroen Geertzen, and Billi Randall (Dec. 2013). "The Centre for Speech, Language and the Brain (CSLB) concept property norms". In: *Behavior research methods* 46 (cit. on p. 35).

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186 (cit. on pp. 23, 24, 28, 32, 100, 102, 134, 137).

Ding, Guiguang, Yuchen Guo, and Jile Zhou (2014). "Collective Matrix Factorization Hashing for Multimodal Data". In: *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pp. 2083–2090 (cit. on p. 55).

Dolan, Bill, Chris Quirk, and Chris Brockett (2004). "Unsupervised Construction of Large Paraphrase Corpora: Exploiting Massively Parallel News Sources". In: *COLING 2004, 20th International Conference on Computational Linguistics,*

Proceedings of the Conference, 23-27 August 2004, Geneva, Switzerland (cit. on pp. 21, 125).

Dong, Li, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon (2019). "Unified language model pre-training for natural language understanding and generation". In: *Advances in Neural Information Processing Systems*, pp. 13042–13054 (cit. on p. 105).

Du, Xinya, Junru Shao, and Claire Cardie (2017). "Learning to Ask: Neural Question Generation for Reading Comprehension". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pp. 1342–1352 (cit. on pp. 40, 109).

Durand, Thibaut (2017). "Weakly supervised learning for visual recognition. (Apprentissage faiblement supervisé pour la reconnaissance visuelle)". PhD thesis. Pierre and Marie Curie University, Paris, France (cit. on p. 26).

Elhoseiny, Mohamed, Yizhe Zhu, Han Zhang, and Ahmed M. Elgammal (2017). "Link the Head to the "Beak": Zero Shot Learning from Noisy Text Description at Part Precision". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 6288–6297 (cit. on p. 97).

Elliott, Desmond (2018). "Adversarial Evaluation of Multimodal Machine Translation". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pp. 2974–2978 (cit. on p. 44).

Engilberge, Martin, Louis Chevallier, Patrick Pérez, and Matthieu Cord (2018). "Finding Beans in Burgers: Deep Semantic-Visual Embedding With Localization". In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 3984–3993 (cit. on p. 50).

Fang, Hao, Saurabh Gupta, Forrest N. Iandola, Rupesh Kumar Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig (2015). "From captions to visual concepts and back". In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 1473–1482 (cit. on p. 50).

Farhadi, A., I. Endres, D. Hoiem, and D. A. Forsyth (2009a). "Describing objects by their attributes". In: *CVPR 2009* (cit. on pp. 46, 69, 89).

Farhadi, A., I. Endres, D. Hoiem, and D. A. Forsyth (2009b). "Describing objects by their attributes". In: *CVPR* (cit. on pp. 46, 47, 58, 66).

Feng, Fangxiang, Xiaojie Wang, and Ruifan Li (2014). "Cross-modal Retrieval with Correspondence Autoencoder". In: *Proceedings of the ACM International Conference on Multimedia, MM '14, Orlando, FL, USA, November 03 - 07, 2014*, pp. 7–16 (cit. on pp. 31, 53, 54).

Feng, Yang, Lin Ma, Wei Liu, and Jiebo Luo (2019). "Unsupervised Image Captioning". In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 4125–4134 (cit. on p. 51).

Feng, Yansong and Mirella Lapata (2010). "Visual Information in Semantic Representation". In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. HLT '10. Los Angeles, California: Association for Computational Linguistics, pp. 91–99 (cit. on p. 36).

Ferrante, M., N. Ferro, and S. Pontarollo (2017). "Are IR Evaluation Measures on an Interval Scale?" In: *SIGIR* (cit. on p. 68).

Ferrari, V. and A. Zisserman (2007). "Learning Visual Attributes". In: *NeurIPS 2007* (cit. on p. 46).

Fincher-Kiefer, Rebecca (Mar. 2001). "Perceptual components of situation models". In: *Memory & Cognition* 29.2, pp. 336–343 (cit. on pp. 34, 119).

Finkelstein, L., E. Gabrilovich, Y. Matias, E. Rivlin, Z. Solan, G. Wolfman, and E. Ruppin (2002). "Placing search in context: the concept revisited". In: *ACM* (cit. on pp. 16, 17, 95).

Firth, J. R. (1957). "A synopsis of linguistic theory 1930-55." In: 1952-59, pp. 1–32 (cit. on p. 14).

Fouhey, David F. and C. Lawrence Zitnick (2014). "Predicting Object Dynamics in Scenes". In: *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pp. 2027–2034 (cit. on p. 136).

Freitag, Markus and Yaser Al-Onaizan (2017). "Beam Search Strategies for Neural Machine Translation". In: *Proceedings of the First Workshop on Neural Machine Translation, NMT@ACL 2017, Vancouver, Canada, August 4, 2017*, pp. 56–60 (cit. on p. 138).

Frome, A., G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M.'A. Ranzato, and T. Mikolov (2013). "DeViSE: A Deep Visual-Semantic Embedding Model". In: *NeurIPS* (cit. on pp. 5, 12, 31, 46–48, 62, 64, 65, 69, 85, 86, 90, 93, 134).

Fu, Y., T. M. Hospedales, T. Xiang, Z.-Y. Fu, and S. Gong (2014). "Transductive Multi-view Embedding for Zero-Shot Recognition and Annotation". In: *ECCV* (cit. on p. 47).

Fu, Y., T. M. Hospedales, T. Xiang, and S. Gong (2015a). "Transductive Multi-View Zero-Shot Learning". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 37.11 (cit. on p. 80).

Fu, Y. and L. Sigal (n.d.). "Semi-supervised Vocabulary-Informed Learning". In: *CVPR 2016* (cit. on p. 47).

Fu, Y., Y. Yang, T. M. Hospedales, T. Xiang, and S. Gong (2015b). "Transductive Multi-label Zero-shot Learning". In: *CoRR* abs/1503.07790. arXiv: 1503.07790 (cit. on pp. 46, 58).

Fuhr, N. (2017). "Some Common Mistakes In IR Evaluation, And How They Can Be Avoided". In: *SIGIR Forum* (cit. on p. 67).

Fujiwara, Y. and G. Irie (2014). "Efficient Label Propagation". In: *ICML* (cit. on p. 81).

Fukui, Akira, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach (2016). "Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pp. 457–468 (cit. on pp. 30, 42).

Fukushima, Kunihiko and Sei Miyake (1982). "Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position". In: *Pattern Recognition* 15.6, pp. 455–469 (cit. on p. 26).

Gao, Haoyuan, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu (2015). "Are You Talking to a Machine? Dataset and Methods for Multilingual Image Question". In: *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 2296–2304 (cit. on p. 42).

Girshick, Ross B., Jeff Donahue, Trevor Darrell, and Jitendra Malik (2014). "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation". In: *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pp. 580–587 (cit. on pp. 25, 52).

Gong, Yunchao, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik (2014). "Improving Image-Sentence Embeddings Using Large Weakly Annotated Photo Collections". In: *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV*, pp. 529–545 (cit. on p. 30).

Goodfellow, I. J., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio (2014). "Generative Adversarial Nets". In: *NeurIPS* (cit. on pp. 6, 44, 51, 88).

Gordon, Jonathan and Benjamin Van Durme (2013). "Reporting Bias and Knowledge Acquisition". In: *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*. AKBC '13. San Francisco, California, USA: ACM, pp. 25–30 (cit. on pp. 3, 34, 35, 50).

Gorti, S. Krishna and J. Ma (2018). "Text-to-Image-to-Text Translation using Cycle Consistent Adversarial Networks". In: *CoRR* abs/1808.04538. arXiv: 1808.04538 (cit. on pp. 6, 12, 51, 52).

Grice, H Paul (1975). "Logic and conversation". In: *1975*, pp. 41–58 (cit. on p. 3).

Grönroos, Stig-Arne, Benoit Huet, Mikko Kurimo, Jorma Laaksonen, Bernard Mérialdo, Phu Pham, Mats Sjöberg, Umut Sulubacak, Jörg Tiedemann, Raphaël Troncy, and Raúl Vázquez (2018). "The MeMAD Submission to the WMT18 Multimodal Translation Task". In: *Proceedings of the Third Conference on Machine Translation: Shared Task Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pp. 603–611 (cit. on p. 44).

Gu, Jiuxiang, Jianfei Cai, Shafiq R. Joty, Li Niu, and Gang Wang (2018). "Look, Imagine and Match: Improving Textual-Visual Cross-Modal Retrieval With Generative Models". In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 7181–7189 (cit. on p. 53).

Hadsell, Raia, Sumit Chopra, and Yann LeCun (2006). "Dimensionality Reduction by Learning an Invariant Mapping". In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA*, pp. 1735–1742 (cit. on p. 31).

Hardoon, David R., Sándor Szedmák, and John Shawe-Taylor (2004). "Canonical Correlation Analysis: An Overview with Application to Learning Methods". In: *Neural Computation* 16.12, pp. 2639–2664 (cit. on p. 53).

Hariharan, Bharath, Pablo Andrés Arbeláez, Ross B. Girshick, and Jitendra Malik (2014). "Simultaneous Detection and Segmentation". In: *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII*, pp. 297–312 (cit. on pp. 2, 25).

Hariharan, Bharath, Pablo Andrés Arbeláez, Ross B. Girshick, and Jitendra Malik (2015). "Hypercolumns for object segmentation and fine-grained localization". In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 447–456 (cit. on p. 25).

Harnad, Stevan (June 1990). "The Symbol Grounding Problem". In: *Phys. D* 42.1-3, pp. 335–346 (cit. on p. 34).

Harris, Z. S (1954). "Distributional structure". In: *Word* (cit. on pp. 13, 118).

He, K., X. Zhang, S. Ren, and J. Sun (2016). "Deep Residual Learning for Image Recognition". In: *CVPR* (cit. on pp. 28, 85, 86).

He, L., X. Xu, H. Lu, Y. Yang, F. Shen, and H. Tao Shen (2017). "Unsupervised cross-modal retrieval through adversarial learning". In: *ICME 2017* (cit. on p. 54).

Helcl, Jindrich, Jindrich Libovický, and Dusan Varis (2018). "CUNI System for the WMT18 Multimodal Translation Task". In: *Proceedings of the Third Conference on Machine Translation: Shared Task Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pp. 616–623 (cit. on p. 44).

Hendricks, Lisa Anne, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach (2018). "Women Also Snowboard: Overcoming Bias in Captioning Models". In: *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part III*, pp. 793–811 (cit. on p. 51).

Hill, F., R. Reichart, and A. Korhonen (2015). "SimLex-999: Evaluating Semantic Models With (Genuine) Similarity Estimation". In: *Computational Linguistics* 41.4 (cit. on pp. 16, 95).

Hill, Felix, Kyunghyun Cho, and Anna Korhonen (2016). "Learning Distributed Representations of Sentences from Unlabelled Data". In: *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational*

Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016, pp. 1367–1377 (cit. on pp. 20, 123, 124).

Hill, Felix and Anna Korhonen (2014a). "Learning Abstract Concept Embeddings from Multi-Modal Data: Since You Probably Can't See What I Mean". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pp. 255–265 (cit. on pp. 36, 37, 128).

Hill, Felix, Roi Reichart, and Anna Korhonen (2014b). "Multi-Modal Models for Concrete and Abstract Concept Meaning". In: *Transactions of the Association for Computational Linguistics* 2, pp. 285–296 (cit. on pp. 30, 36, 38).

Hochreiter, Sepp and Jürgen Schmidhuber (Nov. 1997). "Long Short-Term Memory". In: *Neural Comput.* 9.8, pp. 1735–1780 (cit. on p. 18).

Hu, BingZhang, Yan Gao, Yu Guan, Yang Long, Nicholas Lane, and Thomas Ploetz (2018). "Robust Cross-View Gait Identification with Evidence: A Discriminant Gait GAN (DiGGAN) Approach on 10000 People". In: *CoRR* abs/1811.10493. arXiv: 1811.10493 (cit. on p. 31).

Hu, Minqing and Bing Liu (2004). "Mining and Summarizing Customer Reviews". In: *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD '04. Seattle, WA, USA: ACM, pp. 168–177 (cit. on pp. 21, 125).

Hu, Ronghang, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell (2016). "Natural Language Object Retrieval". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 4555–4564 (cit. on pp. 52, 53).

Huang, Po-Yao, Frederick Liu, Sz-Rung Shiang, Jean Oh, and Chris Dyer (2016). "Attention-based Multimodal Neural Machine Translation". In: *Proceedings of the First Conference on Machine Translation, WMT 2016, colocated with ACL 2016, August 11-12, Berlin, Germany*, pp. 639–645 (cit. on p. 44).

Hudson, Drew A. and Christopher D. Manning (2019). "GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering". In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 6700–6709 (cit. on p. 33).

Jain, Unnat, Svetlana Lazebnik, and Alexander G. Schwing (2018). "Two Can Play This Game: Visual Dialog With Discriminative Question Generation and Answering". In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 5754–5763 (cit. on p. 43).

Jain, Unnat, Ziyu Zhang, and Alexander G. Schwing (2017). "Creativity: Generating Diverse Questions Using Variational Autoencoders". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 5415–5424 (cit. on p. 41).

Jiang, Qing-Yuan and Wu-Jun Li (2017). "Deep Cross-Modal Hashing". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 3270–3278 (cit. on p. 55).

Jiang, Yu, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh (2018). "Pythia v0.1: the Winning Entry to the VQA Challenge 2018". In: *CoRR* abs/1807.09956. arXiv: 1807.09956 (cit. on p. 42).

Johnson, Justin, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick (2017). "CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 1988–1997 (cit. on p. 42).

Johnson, Justin, Andrej Karpathy, and Li Fei-Fei (2016). "DenseCap: Fully Convolutional Localization Networks for Dense Captioning". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 4565–4574 (cit. on pp. 50, 51).

Jones, Karen Spärck (2004). "A statistical interpretation of term specificity and its application in retrieval". In: *Journal of Documentation* 60.5, pp. 493–502 (cit. on p. 22).

Jung, Jaewon and Jongyoul Park (2019). "Visual Relationship Detection with Language prior and Softmax". In: *CoRR* abs/1904.07798. arXiv: 1904.07798 (cit. on p. 49).

Jurie, F., M. Bucher, and S. Herbin (2017). "Generating Visual Representations for Zero-Shot Classification". In: *ICCV* (cit. on pp. 47, 82, 90).

Kalchbrenner, Nal, Edward Grefenstette, and Phil Blunsom (2014). "A Convolutional Neural Network for Modelling Sentences". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pp. 655–665 (cit. on p. 20).

Kampffmeyer, M., Y. Chen, X. Liang, H. Wang, Y. Zhang, and E. P. Xing (2019). "Rethinking Knowledge Graph Propagation for Zero-Shot Learning". In: CVPR (cit. on p. 48).

Karpathy, Andrej and Li Fei-Fei (2017). "Deep Visual-Semantic Alignments for Generating Image Descriptions". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 39.4, pp. 664–676 (cit. on pp. 6, 7, 50, 52, 107, 122).

Khare, V., D. Mahajan, H. Bharadhwaj, V. K. Verma, and P. Rai (2019). "A Generative Framework for Zero-Shot Learning with Adversarial Domain Adaptation". In: *CoRR* abs/1906.03038. arXiv: 1906.03038 (cit. on p. 82).

Kiela, Douwe, Suvrat Bhooshan, Hamed Firooz, and Davide Testuggine (2019). "Supervised Multimodal Bitransformers for Classifying Images and Text". In: *CoRR* abs/1909.02950. arXiv: 1909.02950 (cit. on p. 32).

Kiela, Douwe and Léon Bottou (2014a). "Learning Image Embeddings using Convolutional Neural Networks for Improved Multi-Modal Semantics". In:

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL, pp. 36–45 (cit. on p. 37).

Kiela, Douwe, Alexis Conneau, Allan Jabri, and Maximilian Nickel (2018). "Learning Visually Grounded Sentence Representations". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pp. 408–418 (cit. on pp. 29, 39, 101, 119, 120, 123–125, 129, 130).

Kiela, Douwe, Felix Hill, Anna Korhonen, and Stephen Clark (2014b). "Improving Multi-Modal Representations Using Image Dispersion: Why Less is Sometimes More". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers*, pp. 835–841 (cit. on p. 36).

Kim, Jin-Hwa, Sang-Woo Lee, Dong-Hyun Kwak, Min-Oh Heo, Jeonghee Kim, JungWoo Ha, and Byoung-Tak Zhang (2016). "Multimodal Residual Learning for Visual QA". In: *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 361–369 (cit. on p. 29).

Kim, Jin-Hwa, Kyoung Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang (2017). "Hadamard Product for Low-rank Bilinear Pooling". In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings* (cit. on p. 30).

Kim, Yoon, Carl Denton, Luong Hoang, and Alexander M. Rush (2017). "Structured Attention Networks". In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings* (cit. on p. 19).

Kingma, D. P. and J. Ba (2014a). "Adam: A Method for Stochastic Optimization". In: *CoRR* abs/1412.6980. arXiv: 1412.6980 (cit. on pp. 69, 125).

Kingma, D. P. and M. Welling (2014b). "Auto-Encoding Variational Bayes". In: *ICLR* (cit. on pp. 44, 82).

Kipf, T. N. and M. Welling (2017). "Semi-Supervised Classification with Graph Convolutional Networks". In: *ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings* (cit. on p. 48).

Kiros, Ryan, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler (2015). "Skip-Thought Vectors". In: *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 3294–3302 (cit. on pp. 2, 20, 51, 120, 123–125, 130, 137).

Kodirov, E., T. Xiang, and S. Gong (2017). "Semantic Autoencoder for Zero-Shot Learning". In: *CVPR* (cit. on p. 58).

Kong, Chen, Dahua Lin, Mohit Bansal, Raquel Urtasun, and Sanja Fidler (2014). "What Are You Talking About? Text-to-Image Coreference". In: *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pp. 3558–3565 (cit. on p. 52).

Kottur, Satwik, Ramakrishna Vedantam, José M. F. Moura, and Devi Parikh (2016). "VisualWord2Vec (Vis-W2V): Learning Visually Grounded Word Embeddings Using Abstract Scenes". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 4985–4994 (cit. on pp. 136, 137).

Krishna, R., Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei (2017). "Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations". In: *IJCV* (cit. on pp. 27, 33, 59, 67, 103, 136).

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton (2012). "ImageNet Classification with Deep Convolutional Neural Networks". In: *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*. NIPS'12. Lake Tahoe, Nevada: Curran Associates Inc., pp. 1097–1105 (cit. on pp. 25, 28).

Kumar, Shaishav and Raghavendra Udupa (2011). "Learning Hash Functions for Cross-View Similarity Search". In: *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011*, pp. 1360–1365 (cit. on p. 55).

Lala, Chiraag, Pranava Swaroop Madhyastha, Carolina Scarton, and Lucia Specia (2018). "Sheffield Submissions for WMT18 Multimodal Translation Shared Task". In: *Proceedings of the Third Conference on Machine Translation: Shared Task Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pp. 624–631 (cit. on p. 44).

Lampert, C. H., H. Nickisch, and S. Harmeling (2014). "Attribute-Based Classification for Zero-Shot Visual Object Categorization". In: *TPAMI* (cit. on pp. 47, 58).

Lample, G., A. Conneau, M.'A. Ranzato, L. Denoyer, and H. Jégou (2018). "Word translation without parallel data". In: *ICLR* (cit. on pp. 15, 83, 84, 90).

Lample, Guillaume, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer (2016). "Neural Architectures for Named Entity Recognition". In: *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pp. 260–270 (cit. on p. 13).

Lample, Guillaume, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato (2018a). "Unsupervised Machine Translation Using Monolingual Corpora Only". In: *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings* (cit. on pp. 13, 45).

Lample, Guillaume, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato (2018b). "Phrase-Based & Neural Unsupervised Machine Translation". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pp. 5039–5049 (cit. on p. 45).

Lazaridou, A., N. The Pham, and M. Baroni (2015a). "Combining Language and Vision with a Multimodal Skip-gram Model". In: *NAACL* (cit. on pp. 7, 36, 37, 77, 101).

Lazaridou, A., N. The Pham, and M. Baroni (2015b). "Combining Language and Vision with a Multimodal Skip-gram Model". In: *NAACL* (cit. on p. 93).

Lazaridou, Angeliki, Nghia The Pham, and Marco Baroni (2015). "Combining Language and Vision with a Multimodal Skip-gram Model". In: *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pp. 153–163 (cit. on pp. 12, 119, 120, 124).

Lazaridou, Georgiana Dinu, and Marco Baroni (2015). "Hubness and Pollution: Delving into Cross-Space Mapping for Zero-Shot Learning". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Beijing, China: Association for Computational Linguistics, pp. 270–280 (cit. on p. 130).

Le, Quoc V. and Tomas Mikolov (2014). "Distributed Representations of Sentences and Documents". In: *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pp. 1188–1196 (cit. on pp. 2, 22).

Lecun, Y., L. Bottou, Y. Bengio, and P. Haffner (1998). "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* (cit. on pp. 28, 46, 58).

LeCun, Yann, Yoshua Bengio, and Geoffrey E. Hinton (2015). "Deep learning". In: *Nature* 521.7553, pp. 436–444 (cit. on pp. 1, 12, 53).

LeCun, Yann, Bernhard E. Boser, John S. Denker, Donnie Henderson, Richard E. Howard, Wayne E. Hubbard, and Lawrence D. Jackel (1989). "Backpropagation Applied to Handwritten Zip Code Recognition". In: *Neural Computation* 1.4, pp. 541–551 (cit. on pp. 2, 27).

LeCun, Yann, S. Chopra, and R. Hadsell (2006). "A Tutorial on Energy-Based Learning". In: *Predicting Structured Data* (cit. on p. 62).

Levy, O. and Y. Goldberg (2014). "Neural Word Embedding as Implicit Matrix Factorization". In: *NIPS* (cit. on p. 15).

Li, Bowen, Xiaojuan Qi, Thomas Lukasiewicz, and Philip H. S. Torr (2019). "Controllable Text-to-Image Generation". In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019,*

NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada, pp. 2063–2073 (cit. on p. 137).

Li, Gen, Nan Duan, Yuejian Fang, Daxin Jiang, and Ming Zhou (2019). "Unicoder-VL: A Universal Encoder for Vision and Language by Cross-modal Pre-training". In: *CoRR* abs/1908.06066. arXiv: 1908.06066 (cit. on p. 32).

Li, Liunian Harold, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang (2019). "VisualBERT: A Simple and Performant Baseline for Vision and Language". In: *CoRR* abs/1908.03557. arXiv: 1908.03557 (cit. on pp. 32, 33, 102).

Li, Ruiyu and Jiaya Jia (2016). "Visual Question Answering with Question Representation Update (QRU)". In: *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 4655–4663 (cit. on p. 29).

Li, Yikang, Nan Duan, Bolei Zhou, Xiao Chu, Wanli Ouyang, Xiaogang Wang, and Ming Zhou (2018). "Visual Question Generation as Dual Task of Visual Question Answering". In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 6116–6124 (cit. on pp. 7, 12, 41, 107).

Libovický, Jindrich and Jindrich Helcl (2017). "Attention Strategies for Multi-Source Sequence-to-Sequence Learning". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 2: Short Papers*, pp. 196–202 (cit. on p. 44).

Lin, Chin-Yew (July 2004). "ROUGE: A Package for Automatic Evaluation of Summaries". In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics, pp. 74–81 (cit. on p. 108).

Lin, Tsung-Yi, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick (2014a). "Microsoft COCO: Common Objects in Context". In: *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, pp. 740–755 (cit. on pp. 5, 27, 123).

Lin, Tsung-Yi, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick (2014b). "Microsoft COCO: Common Objects in Context". In: *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, pp. 740–755 (cit. on pp. 33, 89, 107).

Lin, Xiao and Devi Parikh (2015). "Don't just listen, use your imagination: Leveraging visual common sense for non-visual tasks". In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 2984–2993 (cit. on pp. 34, 136).

Lin, Yankai, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu (2015). "Learning Entity and Relation Embeddings for Knowledge Graph Completion". In: *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pp. 2181–2187 (cit. on p. 136).

Lin, Zhouhan, Minwei Feng, Cıcero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio (2017). "A Structured Self-attentive Sentence Embedding". In: *CoRR* abs/1703.03130. arXiv: 1703.03130 (cit. on p. 20).

Lin, Zijia, Guiguang Ding, Mingqing Hu, and Jianmin Wang (2015). "Semantics-preserving hashing for cross-view retrieval". In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 3864–3872 (cit. on p. 55).

Liu, R., Y. Zhao, S. Wei, L. Zheng, and Y. Yang (2019). "Modality-Invariant Image-Text Embedding for Image-Sentence Matching". In: *TOMCCAP* 15.1 (cit. on p. 54).

Liu, Yang and Mirella Lapata (2019). "Text Summarization with Pretrained Encoders". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3721–3731 (cit. on p. 105).

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov (2019). "Roberta: A robustly optimized bert pretraining approach". In: *arXiv preprint arXiv:1907.11692* (cit. on p. 106).

Loeub, Hagar and Roi Reichart (2016). "Effective Combination of Language and Vision Through Model Composition and the R-CCA Method". In: *CoRR* abs/1609.08810. arXiv: 1609.08810 (cit. on pp. 31, 38).

Logeswaran, Lajanugen and Honglak Lee (2018). "An efficient framework for learning sentence representations". In: *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings* (cit. on p. 20).

Long, Jonathan, Evan Shelhamer, and Trevor Darrell (2015). "Fully convolutional networks for semantic segmentation". In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 3431–3440 (cit. on p. 25).

Long, Y., L. Liu, L. Shao, F. Shen, G. Ding, and J. Han (2017). "From Zero-Shot Learning to Conventional Supervised Classification: Unseen Visual Data Synthesis". In: *CVPR* (cit. on p. 58).

Lowe, David G. (2004). "Distinctive Image Features from Scale-Invariant Keypoints". In: *International Journal of Computer Vision* 60.2, pp. 91–110 (cit. on pp. 2, 25).

Lu, Cewu, Ranjay Krishna, Michael S. Bernstein, and Fei-Fei Li (2016a). "Visual Relationship Detection with Language Priors". In: *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*, pp. 852–869 (cit. on p. 49).

Lu, Cewu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei (2016b). "Visual Relationship Detection with Language Priors". In: *European Conference on Computer Vision* (cit. on pp. 12, 49).

Lu, Jiasen, Dhruv Batra, Devi Parikh, and Stefan Lee (2019). "ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks". In: *CoRR* abs/1908.02265. arXiv: 1908.02265 (cit. on pp. 32, 102).

Lu, Jiasen, Anitha Kannan, Jianwei Yang, Devi Parikh, and Dhruv Batra (2017). "Best of Both Worlds: Transferring Knowledge from Discriminative Learning to a Generative Visual Dialog Model". In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pp. 314–324 (cit. on pp. 43, 44).

Lu, Jiasen, Jianwei Yang, Dhruv Batra, and Devi Parikh (2016). "Hierarchical Question-Image Co-Attention for Visual Question Answering". In: *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 289–297 (cit. on p. 42).

Lu, Yongyi, Y.-W. Tai, and C.-K. Tang (2017). "Conditional CycleGAN for Attribute Guided Face Image Generation". In: *CoRR* abs/1705.09966. arXiv: 1705.09966 (cit. on pp. 52, 83).

Lu, Yue M. and Minh N. Do (2007). "Multidimensional Directional Filter Banks and Surfacelets". In: *IEEE Trans. Image Processing* 16.4, pp. 918–931 (cit. on p. 26).

Ma, Lin, Zhengdong Lu, and Hang Li (2016). "Learning to Answer Questions from Image Using Convolutional Neural Network". In: *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pp. 3567–3573 (cit. on p. 42).

Ma, Mingbo, Dapeng Li, Kai Zhao, and Liang Huang (2017). "OSU Multimodal Machine Translation System Report". In: *Proceedings of the Second Conference on Machine Translation, WMT 2017, Copenhagen, Denmark, September 7-8, 2017*, pp. 465–469 (cit. on p. 44).

Maas, Andrew L., Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts (2011). "Learning Word Vectors for Sentiment Analysis". In: *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*, pp. 142–150 (cit. on p. 13).

Maaten, Laurens van der and Geoffrey Hinton (2008). "Visualizing data using t-SNE". In: *Journal of machine learning research* 9.Nov, pp. 2579–2605 (cit. on pp. 81, 128).

Malinowski, Mateusz and Mario Fritz (2014). "A Multi-World Approach to Question Answering about Real-World Scenes based on Uncertain Input". In: *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pp. 1682–1690 (cit. on pp. 42, 101).

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze (2008). *Introduction to information retrieval*. Cambridge University Press (cit. on p. 23).

Mao, Junhua, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy (2016). "Generation and Comprehension of Unambiguous Object Descriptions". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 11–20 (cit. on p. 53).

Marelli, Marco, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli (2014a). "SemEval-2014 Task 1: Evaluation of Compositional Distributional Semantic Models on Full Sentences through Semantic Relatedness and Textual Entailment". In: *Proceedings of the 8th International Workshop on Semantic Evaluation, SemEval@COLING 2014, Dublin, Ireland, August 23-24, 2014*. Pp. 1–8 (cit. on pp. 20, 124).

Marelli, Marco, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli (2014b). "A SICK cure for the evaluation of compositional distributional semantic models". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014*. Pp. 216–223 (cit. on pp. 21, 125).

Massiceti, Daniela, N. Siddharth, Puneet Kumar Dokania, and Philip H. S. Torr (2018). "FlipDial: A Generative Model for Two-Way Visual Dialogue". In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 6097–6105 (cit. on p. 44).

McRae, K., G. S Cree, M. S Seidenberg, and C. McNorgan (2005). "Semantic feature production norms for a large set of living and nonliving things". In: *Behavior research methods* (cit. on pp. 17, 35).

Mensink, T., E. Gavves, and C. Snoek (2014). "COSTA: Co-Occurrence Statistics for Zero-Shot Classification". In: *CVPR* (cit. on p. 59).

Mensink, T., J. J. Verbeek, F. Perronnin, and G. Csurka (2012). "Metric Learning for Large Scale Image Classification: Generalizing to New Classes at Near-Zero Cost". In: *ECCV 2012* (cit. on pp. 46, 58).

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013a). "Efficient Estimation of Word Representations in Vector Space". In: *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings* (cit. on pp. 14, 15).

Mikolov, Tomas, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean (2013b). "Distributed Representations of Words and Phrases and their Compositionality". In: *NeurIPS* (cit. on pp. 2, 14–16, 22, 36, 47, 59, 61, 69, 77, 84, 118, 133).

Mirza, Mehdi and Simon Osindero (2014). "Conditional Generative Adversarial Nets". In: *CoRR* abs/1411.1784. arXiv: 1411.1784 (cit. on p. 51).

Mishra, A., M. S. Krishna Reddy, A. Mittal, and H. A. Murthy (2018). "A Generative Model for Zero Shot Learning Using Conditional Variational Autoencoders". In: *CVPR Workshops* (cit. on pp. 82, 90, 93).

Misra, I., C. L. Zitnick, M. Mitchell, and R. B. Girshick (2016). "Seeing through the Human Reporting Bias: Visual Classifiers from Noisy Human-Centric Labels". In: *CVPR 2016* (cit. on pp. 40, 101).

Moreno, Jose G., Romaric Besançon, Romain Beaumont, Eva D'hondt, Anne-Laure Ligozat, Sophie Rosset, Xavier Tannier, and Brigitte Grau (2017). "Combining Word and Entity Embeddings for Entity Linking". In: *The Semantic Web - 14th International Conference, ESWC 2017, Portorož, Slovenia, May 28 - June 1, 2017, Proceedings, Part I*, pp. 337–352 (cit. on p. 13).

Mostafazadeh, Nasrin, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende (2016). "Generating Natural Questions About an Image". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers* (cit. on pp. 41, 101, 107, 110).

Nakayama, Hideki and Noriki Nishida (2017). "Zero-resource machine translation by multimodal encoder-decoder network with multimedia pivot". In: *Machine Translation* 31.1-2, pp. 49–64 (cit. on p. 45).

Nelson, D. L., C. L. McEvoy, and T. A. Schreiber (Aug. 2004). "The University of South Florida free association, rhyme, and word fragment norms". In: *Behavior Research Methods, Instruments, & Computers* 36.3 (cit. on pp. 17, 95, 128).

Nguyen, Dat Quoc, Dai Quoc Nguyen, Dang Duc Pham, and Son Bao Pham (2016). "A robust transformation-based learning approach using ripple down rules for part-of-speech tagging". In: *AI Commun.* 29.3, pp. 409–422 (cit. on p. 13).

Nilsback, Maria-Elena and Andrew Zisserman (2008). "Automated Flower Classification over a Large Number of Classes". In: *Sixth Indian Conference on Computer Vision, Graphics & Image Processing, ICVGIP 2008, Bhubaneswar, India, 16-19 December 2008*, pp. 722–729 (cit. on p. 52).

Nitish, Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov (2014). "Dropout: a simple way to prevent neural networks from overfitting". In: *J. Mach. Learn. Res.* 15.1, pp. 1929–1958 (cit. on p. 27).

Noh, Hyeonwoo, Paul Hongsuck Seo, and Bohyung Han (2016). "Image Question Answering Using Convolutional Neural Network with Dynamic Parameter Prediction". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 30–38 (cit. on p. 42).

Norman, Donald A (1972). "Memory, knowledge, and the answering of questions." In: (cit. on p. 118).

Norouzi, M., T.Mikolov, S. Bengio, Y. Singer, J. Shlens, A. Frome, G. Corrado, and J. Dean (2014). "Zero-Shot Learning by Convex Combination of Semantic Embeddings". In: *ICLR*. arXiv: 1312.5650 (cit. on pp. 7, 46–48, 83, 85, 86, 90, 93, 96).

Novikova, Jekaterina, Ondrej Dusek, Amanda Cercas Curry, and Verena Rieser (2017). "Why We Need New Evaluation Metrics for NLG". In: *Proceedings of*

the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017, pp. 2241–2252 (cit. on p. 51).

Oord, Aäron van den, Nal Kalchbrenner, Lasse Espeholt, Koray Kavukcuoglu, Oriol Vinyals, and Alex Graves (2016). "Conditional Image Generation with PixelCNN Decoders". In: *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 4790–4798 (cit. on p. 25).

Pang, Bo and Lillian Lee (2004). "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts". In: *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, 21-26 July, 2004, Barcelona, Spain.* Pp. 271–278 (cit. on pp. 21, 125).

Pang, Bo and Lillian Lee (2005). "Seeing Stars: Exploiting Class Relationships for Sentiment Categorization with Respect to Rating Scales". In: *ACL 2005, 43rd Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 25-30 June 2005, University of Michigan, USA*, pp. 115–124 (cit. on pp. 21, 125).

Pang, Bo and Lillian Lee (2007). "Opinion Mining and Sentiment Analysis". In: *Foundations and Trends in Information Retrieval* 2.1-2, pp. 1–135 (cit. on pp. 1, 5, 13).

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2002). "BLEU: a method for automatic evaluation of machine translation". In: *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, pp. 311–318 (cit. on pp. 51, 108).

Parikh, D. and K. Grauman (2011). "Relative attributes". In: *ICCV* (cit. on p. 46).

Patro, Badri N., Sandeep Kumar, Vinod Kumar Kurmi, and Vinay P. Namboodiri (2018a). "Multimodal Differential Network for Visual Question Generation". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pp. 4002–4012 (cit. on pp. 29, 40, 41).

Patro, Badri N., Vinod K. Kurmi, Sandeep Kumar, and Vinay P. Namboodiri (2020). *Deep Bayesian Network for Visual Question Generation*. arXiv: 2001.08779 [cs.CV] (cit. on pp. 40, 41, 108).

Patro, Badri N. and Vinay P. Namboodiri (2018b). "Differential Attention for Visual Question Answering". In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 7680–7688 (cit. on p. 42).

Patro, Badri N. and Vinay P. Namboodiri (2019). "Deep Exemplar Networks for VQA and VQG". In: *CoRR* abs/1912.09551. arXiv: 1912.09551 (cit. on pp. 40, 41, 107, 108).

Patterson, G. and J. Hays (2012). "SUN Attribute Database: Discovering, Annotating, and Recognizing Scene Attributes". In: *CVPR 2012* (cit. on p. 47).

Pelletier, Francis (Mar. 2001). "Did Frege Believe Frege's Principle?" In: *Journal of Logic, Language and Information* 10, pp. 87–114 (cit. on p. 58).

Peng, Yuxin, Xin Huang, and Jinwei Qi (2016). "Cross-Media Shared Representation by Hierarchical Learning with Multiple Deep Networks". In: *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, pp. 3846–3853 (cit. on p. 53).

Peng, Yuxin, Jinwei Qi, Xin Huang, and Yuxin Yuan (2018). "CCL: Cross-modal Correlation Learning With Multigrained Fusion by Hierarchical Network". In: *IEEE Trans. Multimedia* 20.2, pp. 405–420 (cit. on p. 53).

Pennington, J., R. Socher, and C. D. Manning (2014). "Glove: Global Vectors for Word Representation". In: *EMNLP* (cit. on pp. 14, 15, 47).

Pereira, Jose Costa, Emanuele Coviello, Gabriel Doyle, Nikhil Rasiwasia, Gert R. G. Lanckriet, Roger Levy, and Nuno Vasconcelos (2014). "On the Role of Correlation and Abstraction in Cross-Modal Multimedia Retrieval". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 36.3, pp. 521–535 (cit. on p. 54).

Peters, Matthew E., Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer (2018). "Deep Contextualized Word Representations". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pp. 2227–2237 (cit. on pp. 15, 118).

Petrov, Slav, Dipanjan Das, and Ryan T. McDonald (2012). "A Universal Part-of-Speech Tagset". In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, May 23-25, 2012*, pp. 2089–2096 (cit. on p. 13).

Pires, Telmo, Eva Schlinger, and Dan Garrette (2019). "How Multilingual is Multilingual BERT?" In: *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pp. 4996–5001 (cit. on p. 102).

Plummer, Bryan A., Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik (2015). "Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models". In: *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pp. 2641–2649 (cit. on pp. 53, 128).

Plummer, Bryan A., Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik (2017). "Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models". In: *International Journal of Computer Vision* 123.1, pp. 74–93 (cit. on p. 52).

Pontiki, Maria, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammad Al-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, Véronique Hoste, Marianna Apidianaki, Xavier Tannier, Natalia V. Loukachevitch, Evgeniy V. Kotelnikov, Núria Bel,

Salud Marıa Jiménez Zafra, and Gülsen Eryigit (2016). "SemEval-2016 Task 5: Aspect Based Sentiment Analysis". In: *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*, pp. 19–30 (cit. on p. 13).

Pulvermüller, Friedemann (Aug. 2005). "Brain Mechanisms Linking Language and Action". In: *Nature reviews. Neuroscience* 6, pp. 576–82 (cit. on p. 35).

Qi, Jinwei and Yuxin Peng (2018). "Cross-modal Bidirectional Translation via Reinforcement Learning". In: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pp. 2630–2636 (cit. on p. 53).

Qiu, Guoping (2002). "Indexing chromatic and achromatic patterns for content-based colour image retrieval". In: *Pattern Recognition* 35.8, pp. 1675–1686 (cit. on p. 26).

Radford, Alec, Luke Metz, and Soumith Chintala (2016). "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks". In: *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings* (cit. on pp. 25, 51).

Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever (2018). "Improving language understanding by generative pre-training". In: (cit. on pp. 23, 110).

Rajpurkar, Pranav, Jian Zhang, Konstantin Lopyrev, and Percy Liang (2016). "SQuAD: 100, 000+ Questions for Machine Comprehension of Text". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pp. 2383–2392 (cit. on pp. 17, 40).

Ramakrishnan, Sainandan, Aishwarya Agrawal, and Stefan Lee (2018). "Overcoming Language Priors in Visual Question Answering with Adversarial Regularization". In: *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pp. 1548–1558 (cit. on p. 43).

Ranzato, Marc'Aurelio, Sumit Chopra, Michael Auli, and Wojciech Zaremba (2016). "Sequence Level Training with Recurrent Neural Networks". In: *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings* (cit. on p. 51).

Real, E., A. Aggarwal, Y. Huang, and Q. V. Le (2019). "Regularized Evolution for Image Classifier Architecture Search". In: *AAAI* (cit. on pp. 46, 58).

Redmon, Joseph, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi (2016). "You Only Look Once: Unified, Real-Time Object Detection". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 779–788 (cit. on p. 25).

Reed, Scott E., Zeynep Akata, Honglak Lee, and Bernt Schiele (2016a). "Learning Deep Representations of Fine-Grained Visual Descriptions". In: *2016 IEEE*

Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, pp. 49–58 (cit. on p. 51).

Reed, Scott E., Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee (2016b). "Generative Adversarial Text to Image Synthesis". In: *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pp. 1060–1069 (cit. on p. 51).

Reichart, Roi and Anna Korhonen (2013). "Improved Lexical Acquisition through DPP-based Verb Clustering". In: *ACL* (cit. on p. 38).

Reif, Emily, Ann Yuan, Martin Wattenberg, Fernanda B. Viégas, Andy Coenen, Adam Pearce, and Been Kim (2019). "Visualizing and Measuring the Geometry of BERT". In: *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pp. 8592–8600 (cit. on p. 112).

Ren, Mengye, Ryan Kiros, and Richard S. Zemel (2015). "Exploring Models and Data for Image Question Answering". In: *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pp. 2953–2961 (cit. on p. 42).

Ren, Shaoqing, Kaiming He, Ross B. Girshick, and Jian Sun (2017). "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 39.6, pp. 1137–1149 (cit. on pp. 25, 32, 42, 52, 78, 103, 136).

Ren, Zhou, Xiaoyu Wang, Ning Zhang, Xutao Lv, and Li-Jia Li (2017). "Deep Reinforcement Learning-Based Image Captioning with Embedding Reward". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 1151–1159 (cit. on p. 51).

Rohrbach, Anna, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele (Oct. 2016). "Grounding of textual phrases in images by reconstruction". In: *European Conference on Computer Vision (ECCV)*. Oral. Springer. Amsterdam, The Netherlands: Springer (cit. on p. 53).

Roller, Stephen and Sabine Schulte im Walde (Oct. 2013). "A Multimodal LDA Model Integrating Textual, Cognitive and Visual Modalities". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013)*. Seattle, WA, pp. 1146–1157 (cit. on p. 36).

Romera-Paredes, B. and P. H. S. Torr (2015). "An embarrassingly simple approach to zero-shot learning". In: *ICML* (cit. on p. 47).

Rus, Vasile, Brendan Wyse, Paul Piwek, Mihai C. Lintean, Svetlana Stoyanchev, and Cristian Moldovan (2010). "The First Question Generation Shared Task Evaluation Challenge". In: *INLG 2010 - Proceedings of the Sixth International Natural Language Generation Conference, July 7-9, 2010, Trim, Co. Meath, Ireland* (cit. on p. 40).

Rush, Alexander M., Sumit Chopra, and Jason Weston (2015). "A Neural Attention Model for Abstractive Sentence Summarization". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pp. 379–389 (cit. on pp. 1, 13).

Sadeghi, Fereshteh, Santosh K. Divvala, and Ali Farhadi (June 2015). *VisKE: Visual knowledge extraction and question answering by visual verification of relation phrases* (cit. on pp. 49, 136).

Sadeghi, Mohammad Amin and Ali Farhadi (2011). "Recognition using visual phrases". In: *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011*, pp. 1745–1752 (cit. on p. 49).

Salton, Gerard and Michael McGill (1984). *Introduction to Modern Information Retrieval*. McGraw-Hill Book Company (cit. on p. 22).

Sang, Erik F. Tjong Kim and Fien De Meulder (2003). "Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition". In: *Proceedings of the Seventh Conference on Natural Language Learning, CoNLL 2003, Held in cooperation with HLT-NAACL 2003, Edmonton, Canada, May 31 - June 1, 2003*, pp. 142–147 (cit. on p. 17).

Schakel, A. M. J. and B. J. Wilson (2015). "Measuring Word Significance using Distributed Representations of Words". In: *CoRR* abs/1508.02297. arXiv: 1508.02297 (cit. on p. 71).

Schuster, Mike and Kuldip K. Paliwal (1997). "Bidirectional recurrent neural networks". In: *IEEE Trans. Signal Process.* 45.11, pp. 2673–2681 (cit. on p. 19).

Scialom, Thomas, Benjamin Piwowarski, and Jacopo Staiano (2019). "Self-Attention Architectures for Answer-Agnostic Neural Question Generation". In: *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pp. 6027–6032 (cit. on p. 40).

Sermanet, Pierre, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun (2014). "OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks". In: *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings* (cit. on pp. 2, 25).

Sharma, Abhishek, Abhishek Kumar, Hal Daumé III, and David W. Jacobs (2012). "Generalized Multiview Analysis: A discriminative latent space". In: *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, pp. 2160–2167 (cit. on p. 53).

Sharma, Piyush, Nan Ding, Sebastian Goodman, and Radu Soricut (2018). "Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pp. 2556–2565 (cit. on p. 33).

Shih, Kevin J., Saurabh Singh, and Derek Hoiem (2016). "Where to Look: Focus Regions for Visual Question Answering". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 4613–4621 (cit. on p. 42).

Silberer, Carina, Vittorio Ferrari, and Mirella Lapata (2013). "Models of Semantic Representation with Visual Attributes". In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, pp. 572–582 (cit. on pp. 30, 35–38).

Silberer, Carina and Mirella Lapata (2012). "Grounded Models of Semantic Representation". In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. EMNLP-CoNLL '12. Jeju Island, Korea: Association for Computational Linguistics, pp. 1423–1433 (cit. on pp. 6, 30, 35, 38).

Silberer, Carina and Mirella Lapata (2014). "Learning Grounded Meaning Representations with Autoencoders". In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 1: Long Papers*, pp. 721–732 (cit. on pp. 16, 36, 95, 119, 120).

Simonyan, Karen and Andrew Zisserman (2014). "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *CoRR* abs/1409.1556. arXiv: 1409.1556 (cit. on pp. 26, 28).

Smith, Samuel L., David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla (2017). "Offline bilingual word vectors, orthogonal transformations and the inverted softmax". In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings* (cit. on p. 15).

Socher, Richard, Andrej Karpathy, Quoc V. Le, Christopher D. Manning, and Andrew Y. Ng (2014a). "Grounded Compositional Semantics for Finding and Describing Images with Sentences". In: *Transactions of the Association of Computational Linguistics* 2, pp. 207–218 (cit. on pp. 6, 32, 50).

Socher, Richard, Andrej Karpathy, Quoc V. Le, Christopher D. Manning, and Andrew Y. Ng (2014b). "Grounded Compositional Semantics for Finding and Describing Images with Sentences". In: *TACL* 2, pp. 207–218 (cit. on p. 101).

Socher, Richard, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts (2013). "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank". In: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pp. 1631–1642 (cit. on pp. 20, 21, 125).

Song, Linfeng, Zhiguo Wang, and Wael Hamza (2017). "A Unified Query-based Generative Model for Question Generation and Question Answering". In: *CoRR* abs/1709.01058. arXiv: 1709.01058 (cit. on p. 40).

Sordoni, Alessandro, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan (2015). "A Neural Network Approach to Context-Sensitive Generation of Conversational Responses". In: *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*. Ed. by Rada Mihalcea, Joyce Yue Chai, and Anoop Sarkar. The Association for Computational Linguistics, pp. 196–205 (cit. on p. 13).

Strub, Florian, Harm de Vries, Jérémie Mary, Bilal Piot, Aaron C. Courville, and Olivier Pietquin (2017). "End-to-end optimization of goal-driven and visually grounded dialogue systems". In: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pp. 2765–2771 (cit. on p. 43).

Su, Weijie, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai (2019). "VL-BERT: Pre-training of Generic Visual-Linguistic Representations". In: *CoRR* abs/1908.08530. arXiv: 1908.08530 (cit. on pp. 32, 102).

Su, Yuanhang, Kai Fan, Nguyen Bach, C.-C. Jay Kuo, and Fei Huang (2019). "Unsupervised Multi-Modal Neural Machine Translation". In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 10482–10491 (cit. on p. 45).

Sun, Chen, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid (2019). "VideoBERT: A Joint Model for Video and Language Representation Learning". In: *CoRR* abs/1904.01766. eprint: 1904.01766 (cit. on pp. 32, 131).

Sun, Yu, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang (2019). "Ernie 2.0: A continual pre-training framework for language understanding". In: *arXiv preprint arXiv:1907.12412* (cit. on p. 106).

Sutskever, Ilya, Oriol Vinyals, and Quoc V Le (2014). "Sequence to sequence learning with neural networks". In: *Advances in neural information processing systems*, pp. 3104–3112 (cit. on pp. 6, 19, 43, 50, 105).

Szegedy, Christian, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich (2015). "Going deeper with convolutions". In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 1–9 (cit. on p. 28).

Szegedy, Christian, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna (2016a). "Rethinking the Inception Architecture for Computer Vision". In: *CVPR* (cit. on pp. 25, 69, 90).

Szegedy, Christian, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna (2016b). "Rethinking the Inception Architecture for Computer Vision". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 2818–2826 (cit. on p. 125).

Tan, Hao and Mohit Bansal (2019). "LXMERT: Learning Cross-Modality Encoder Representations from Transformers". In: *CoRR* abs/1908.07490. arXiv: 1908.07490 (cit. on pp. 32, 33, 102, 103).

Teney, Damien, Peter Anderson, Xiaodong He, and Anton van den Hengel (2018). "Tips and Tricks for Visual Question Answering: Learnings From the 2017 Challenge". In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 4223–4232 (cit. on p. 42).

Teney, Damien, Lingqiao Liu, and Anton van den Hengel (2016). "Graph-Structured Representations for Visual Question Answering". In: *CoRR* abs/1609.05600. arXiv: 1609.05600 (cit. on pp. 33, 41, 107).

Tenney, Ian, Dipanjan Das, and Ellie Pavlick (2019). "BERT Rediscovers the Classical NLP Pipeline". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4593–4601 (cit. on p. 112).

Therriault, David, Richard Yaxley, and Rolf Zwaan (June 2009). "The role of color diagnosticity in object recognition and representation". In: *Cognitive processing* 10, pp. 335–42 (cit. on p. 35).

Tong, Simon and Daphne Koller (2001). "Support Vector Machine Active Learning with Applications to Text Classification". In: *J. Mach. Learn. Res.* 2, pp. 45–66 (cit. on pp. 2, 25).

Torralba, A., K. P. Murphy, and W. T. Freeman (2010). "Using the forest to see the trees: exploiting context for visual object detection and localization". In: *ACM* (cit. on p. 69).

Uijlings, Jasper R. R., Koen E. A. van de Sande, Theo Gevers, and Arnold W. M. Smeulders (2013). "Selective Search for Object Recognition". In: *International Journal of Computer Vision* 104.2, pp. 154–171 (cit. on pp. 25, 45, 52).

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pp. 5998–6008 (cit. on pp. 19, 102, 105, 106).

Vedantam, Ramakrishna, C Lawrence Zitnick, and Devi Parikh (2015a). "Cider: Consensus-based image description evaluation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575 (cit. on p. 108).

Vedantam, Ramakrishna, Xiao Lin, Tanmay Batra, C. Lawrence Zitnick, and Devi Parikh (2015b). "Learning Common Sense through Visual Abstraction". In: *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pp. 2542–2550 (cit. on p. 136).

Verma, V. Kumar, G. Arora, A. Mishra, and P. Rai (2018). "Generalized Zero-Shot Learning via Synthesized Examples". In: *CVPR* (cit. on pp. 90, 93).

Verma, V. Kumar and P. Rai (2017). "A Simple Exponential Family Framework for Zero-Shot Learning". In: *ECML PKDD 2017* (cit. on p. 81).

Vilnis, Luke and Andrew McCallum (2014). "Word Representations via Gaussian Embedding". In: *CoRR* abs/1412.6623. arXiv: 1412.6623 (cit. on p. 15).

Vinyals, Oriol and Quoc V. Le (2015a). "A Neural Conversational Model". In: *CoRR* abs/1506.05869. arXiv: 1506.05869 (cit. on p. 13).

Vinyals, Oriol, Alexander Toshev, Samy Bengio, and Dumitru Erhan (2015b). "Show and tell: A neural image caption generator". In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 3156–3164 (cit. on p. 50).

Vries, Harm de, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron C. Courville (2017). "GuessWhat?! Visual Object Discovery through Multi-modal Dialogue". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 4466–4475 (cit. on p. 43).

Vulic, Ivan and Marie-Francine Moens (2015). "Monolingual and Cross-Lingual Information Retrieval Models Based on (Bilingual) Word Embeddings". In: *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, Santiago, Chile, August 9-13, 2015*, pp. 363–372 (cit. on p. 22).

W. Barsalou, Lawrence (Sept. 1999). "Perceptual Symbol Systems". In: 22, 577–609, discussion 610 (cit. on p. 34).

Wah, Catherine, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie (2011). "The caltech-ucsd birds-200-2011 dataset". In: (cit. on p. 66).

Wan, Ziyu, D. Chen, Y. Li, X. Yan, J. Zhang, Y. Yu, and J. Liao (2019). "Transductive Zero-Shot Learning with Visual Structure Constraint". In: *NeurIPS* (cit. on pp. 80–82).

Wang, Bokun, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen (2017). "Adversarial Cross-Modal Retrieval". In: *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*, pp. 154–162 (cit. on p. 54).

Wang, Hong, Wenhan Xiong, Mo Yu, Xiaoxiao Guo, Shiyu Chang, and William Yang Wang (2019). "Sentence Embedding Alignment for Lifelong Relation Extraction". In: *NAACL*. arXiv: 1903.02588 (cit. on p. 118).

Wang, Jian, Yonghao He, Cuicui Kang, Shiming Xiang, and Chunhong Pan (2015). "Image-Text Cross-Modal Retrieval via Modality-Specific Feature Learning". In: *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, Shanghai, China, June 23-26, 2015*, pp. 347–354 (cit. on pp. 53, 54).

Wang, Kaiye, Qiyue Yin, Wei Wang, Shu Wu, and Liang Wang (2016). "A Comprehensive Survey on Cross-modal Retrieval". In: *CoRR* abs/1607.06215. arXiv: 1607.06215 (cit. on p. 53).

Wang, Liwei, Yin Li, and Svetlana Lazebnik (2016). "Learning Deep Structure-Preserving Image-Text Embeddings". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 5005–5013 (cit. on pp. 53, 122).

Wang, Wei, Xiaoyan Yang, Beng Chin Ooi, Dongxiang Zhang, and Yueting Zhuang (2016). "Effective deep learning-based multi-modal retrieval". In: *VLDB J.* 25.1, pp. 79–101 (cit. on p. 54).

Wang, Weiran, Raman Arora, Karen Livescu, and Jeff A. Bilmes (2015). "On Deep Multi-View Representation Learning". In: *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pp. 1083–1092 (cit. on p. 53).

Wang, Weiran and Karen Livescu (2016). "Large-Scale Approximate Kernel Canonical Correlation Analysis". In: *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings* (cit. on p. 53).

Wang, Wenlin, Y. Pu, V. Kumar Verma, K. Fan, Y. Zhang, C. Chen, P. Rai, and L. Carin (2018). "Zero-Shot Learning via Class-Conditioned Deep Generative Models". In: *AAAI* (cit. on pp. 82, 90, 93).

Wang, X., Y. Ye, and A. Gupta (2018). "Zero-Shot Recognition via Semantic Embeddings and Knowledge Graphs". In: *CVPR 2018* (cit. on p. 48).

Wang, Yu-Siang, Hung-Ting Su, Chen-Hsi Chang, and Winston H. Hsu (2019). "Video Question Generation via Cross-Modal Self-Attention Networks Learning". In: *CoRR* abs/1907.03049. arXiv: 1907.03049 (cit. on p. 116).

Wang, Zhen, Jianwen Zhang, Jianlin Feng, and Zheng Chen (2014). "Knowledge Graph Embedding by Translating on Hyperplanes". In: *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada*, pp. 1112–1119 (cit. on p. 136).

Wei, Yunchao, Yao Zhao, Canyi Lu, Shikui Wei, Luoqi Liu, Zhenfeng Zhu, and Shuicheng Yan (2017). "Cross-Modal Retrieval With CNN Visual Features: A New Baseline". In: *IEEE Trans. Cybernetics* 47.2, pp. 449–460 (cit. on p. 53).

Welinder, P., S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona (2010). *Caltech-UCSD Birds 200*. Tech. rep. CNS-TR-2010-001. California Institute of Technology (cit. on pp. 47, 52).

Weston, Jason, S. Bengio, and N. Usunier (2010). "Large scale image annotation: learning to rank with joint word-image embeddings". In: *Machine Learning* (cit. on p. 30).

Weston, Jason, S. Bengio, and N. Usunier (2011). "WSABIE: Scaling Up to Large Vocabulary Image Annotation". In: *IJCAI* (cit. on p. 64).

Weston, Jason, Sumit Chopra, and Antoine Bordes (2015). "Memory Networks". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings* (cit. on pp. 5, 13).

Wiebe, Janyce and Claire Cardie (2005). "Annotating expressions of opinions and emotions in language. Language Resources and Evaluation". In: *Language Resources and Evaluation (formerly Computers and the Humanities*, p. 2005 (cit. on pp. 21, 125).

Winograd, Terry (1971). "Procedures as a representation for data in a computer program for understanding natural language". PhD thesis. MIT (cit. on p. 2).

Wiseman, Sam and Alexander M. Rush (2016). "Sequence-to-Sequence Learning as Beam-Search Optimization". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pp. 1296–1306 (cit. on p. 138).

Wittgenstein, L. (1922). "Tractatus Logico-Philosophicus". In: *London: Routledge, 1981*. Ed. by D.F.Pears (cit. on p. 3).

Wu, Chenfei, Jinlai Liu, Xiaojie Wang, and Ruifan Li (2019). "Differential Networks for Visual Question Answering". In: *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pp. 8997–9004 (cit. on p. 42).

Wu, Jianlong, Zhouchen Lin, and Hongbin Zha (2017). "Joint Latent Subspace Learning and Regression for Cross-Modal Retrieval". In: *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, pp. 917–920 (cit. on p. 30).

Wu, Qi, Peng Wang, Chunhua Shen, Ian D. Reid, and Anton van den Hengel (2018). "Are You Talking to Me? Reasoned Visual Dialog Generation Through Adversarial Learning". In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 6106–6115 (cit. on pp. 43, 44).

Wu, Shijie and Mark Dredze (2019). "Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT". In: *CoRR* abs/1904.09077. arXiv: 1904.09077 (cit. on p. 102).

Xian, Yongqin, Z. Akata, G. Sharma, Q. N. Nguyen, M. Hein, and B. Schiele (2016). "Latent Embeddings for Zero-Shot Classification". In: *CVPR* (cit. on p. 48).

Xian, Yongqin, T. Lorenz, B. Schiele, and Z. Akata (2018). "Feature Generating Networks for Zero-Shot Learning". In: *CVPR 2018* (cit. on p. 82).

Xiao, Fanyi, Leonid Sigal, and Yong Jae Lee (2017). "Weakly-Supervised Visual Grounding of Phrases with Linguistic Structures". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 5253–5262 (cit. on p. 119).

Xing, Eric P., Andrew Y. Ng, Michael I. Jordan, and Stuart J. Russell (2002). "Distance Metric Learning with Application to Clustering with Side-Information". In: *Advances in Neural Information Processing Systems 15 [Neural Information*

Processing Systems, NIPS 2002, December 9-14, 2002, Vancouver, British Columbia, Canada], pp. 505–512 (cit. on p. 31).

Xu, Huijuan and Kate Saenko (2016). "Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering". In: *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VII*, pp. 451–466 (cit. on pp. 32, 40, 42).

Xu, Kelvin, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio (2015). "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention". In: *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pp. 2048–2057 (cit. on pp. 32, 50).

Yan, Xinchen, Jimei Yang, Kihyuk Sohn, and Honglak Lee (2016). "Attribute2Image: Conditional Image Generation from Visual Attributes". In: *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV*, pp. 776–791 (cit. on p. 50).

Yang, Yezhou, Yi Li, Cornelia Fermüller, and Yiannis Aloimonos (2015). "Neural Self Talk: Image Understanding via Continuous Questioning and Answering". In: *CoRR* abs/1512.03460. arXiv: 1512.03460 (cit. on pp. 40, 41, 108).

Yang, Zichao, Xiaodong He, Jianfeng Gao, Li Deng, and Alexander J. Smola (2016). "Stacked Attention Networks for Image Question Answering". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 21–29 (cit. on p. 42).

Yatskar, Mark, Vicente Ordonez, and Ali Farhadi (2016). "Stating the Obvious: Extracting Visual Common Sense Knowledge". In: *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pp. 193–198 (cit. on p. 34).

Ye, M. and Y. Guo (2017). "Zero-Shot Classification with Discriminative Semantic Representation Learning". In: *CVPR 2017* (cit. on pp. 81, 82).

Yongqin Xian C. H. Lampert, B. Schiele and Z. Akata (2019). "Zero-Shot Learning - A Comprehensive Evaluation of the Good, the Bad and the Ugly". In: *IEEE* (cit. on pp. 47, 81, 91).

Young, Peter, Alice Lai, Micah Hodosh, and Julia Hockenmaier (2014). "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions". In: *TACL* 2, pp. 67–78 (cit. on p. 53).

Yu, Ruichi, Ang Li, Vlad I. Morariu, and Larry S. Davis (2017). "Visual Relationship Detection with Internal and External Linguistic Knowledge Distillation". In: *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 1068–1076 (cit. on p. 49).

Yu, Zhou, Jun Yu, Jianping Fan, and Dacheng Tao (2017). "Multi-modal Factorized Bilinear Pooling with Co-attention Learning for Visual Question Answering".

In: *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 1839–1848 (cit. on p. 30).

Zablocki, E., P. Bordes, L. Soulier, B. Piwowarski, and P. Gallinari (2019). "Context-Aware Zero-Shot Learning for Object Recognition". In: *ICML* (cit. on pp. 46, 89, 101).

Zablocki, Eloi, Benjamin Piwowarski, Laure Soulier, and Patrick Gallinari (2018a). "Learning Multi-Modal Word Representation Grounded in Visual Context". In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pp. 5626–5633 (cit. on p. 77).

Zablocki, Eloi, Benjamin Piwowarski, Laure Soulier, and Patrick Gallinari (2018b). "Learning Multi-Modal Word Representation Grounded in Visual Context". In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018* (cit. on pp. 120, 129).

Zadeh, Amir, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency (2016). "MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos". In: *CoRR* abs/1606.06259. arXiv: 1606.06259 (cit. on p. 12).

Zeiler, Matthew D. and Rob Fergus (2014). "Visualizing and Understanding Convolutional Networks". In: *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, pp. 818–833 (cit. on p. 27).

Zhai, Deming, Hong Chang, Shiguang Shan, Xilin Chen, and Wen Gao (2012). "Multiview Metric Learning with Global Consistency and Local Smoothness". In: *ACM TIST* 3.3, 53:1–53:22 (cit. on pp. 53, 54).

Zhai, Xiaohua, Yuxin Peng, and Jianguo Xiao (2013). "Heterogeneous Metric Learning with Joint Graph Regularization for Cross-Media Retrieval". In: *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, July 14-18, 2013, Bellevue, Washington, USA* (cit. on pp. 53, 54).

Zhang, Dongqing and Wu-Jun Li (2014). "Large-Scale Supervised Multimodal Hashing with Semantic Correlation Maximization". In: *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada*, pp. 2177–2183 (cit. on p. 55).

Zhang, Han, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. N. Metaxas (2019). "StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 41.8 (cit. on p. 51).

Zhang, Han, Tao Xu, and Hongsheng Li (2017). "StackGAN: Text to Photo-Realistic Image Synthesis with Stacked Generative Adversarial Networks". In: *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pp. 5908–5916 (cit. on p. 51).

Zhang, Richard, Phillip Isola, and Alexei A. Efros (2016). "Colorful Image Colorization". In: *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III*, pp. 649–666 (cit. on p. 25).

Zhang, Shijie, Lizhen Qu, Shaodi You, Zhenglu Yang, and Jiawan Zhang (2017). "Automatic Generation of Grounded Visual Questions". In: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pp. 4235–4243 (cit. on p. 41).

Zhang, Z. and V. Saligrama (2015). "Zero-Shot Learning via Semantic Similarity Embedding". In: *ICCV 2015* (cit. on p. 47).

Zhao, A., M. Ding, J. Guan, Z. Lu, T. Xiang, and J.-R. Wen (2018). "Domain-Invariant Projection Learning for Zero-Shot Recognition". In: *NeurIPS* (cit. on pp. 81, 82, 89, 90, 93).

Zhao, B., B. Chang, Z. Jie, and L. Sigal (2018). "Modular Generative Adversarial Networks". In: *CoRR* abs/1804.03343. arXiv: 1804.03343 (cit. on p. 66).

Zhao, J., J. Zhang, Z. Li, J.-N. Hwang, Y. Gao, Z. Fang, X. Jiang, and B. Huang (2019). "DD-CycleGAN: Unpaired image dehazing via Double-Discriminator Cycle-Consistent Generative Adversarial Network". In: *EAAI* (cit. on pp. 52, 83).

Zhao, Yao, Xiaochuan Ni, Yuanyuan Ding, and Qifa Ke (2018). "Paragraph-level Neural Question Generation with Maxout Pointer and Gated Self-attention Networks". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pp. 3901–3910 (cit. on p. 40).

Zhou, Bolei, Àgata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba (2018). "Places: A 10 Million Image Database for Scene Recognition". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 40.6, pp. 1452–1464 (cit. on p. 41).

Zhou, D., O. Bousquet, T. Navin Lal, J. Weston, and B. Schölkopf (2003). "Learning with Local and Global Consistency". In: *NeurIPS 2003* (cit. on p. 80).

Zhou, Luowei, Nathan Louis, and Jason J Corso (2018). "Weakly-supervised video object grounding from text by loss weighting and object interaction". In: *arXiv preprint arXiv:1805.02834* (cit. on p. 32).

Zhou, Qingyu, Nan Yang, Furu Wei, Chuanqi Tan, Hangbo Bao, and Ming Zhou (2017). "Neural Question Generation from Text: A Preliminary Study". In: *Natural Language Processing and Chinese Computing - 6th CCF International Conference, NLPCC 2017, Dalian, China, November 8-12, 2017, Proceedings*, pp. 662–671 (cit. on p. 40).

Zhu, J.-Yan, T. Park, P. Isola, and A. A. Efros (2017). "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks". In: *ICCV* (cit. on pp. 25, 51, 52, 55, 83, 87, 134).

Zhu, Yizhe, Mohamed Elhoseiny, Bingchen Liu, Xi Peng, and Ahmed Elgammal (2018). "A Generative Adversarial Approach for Zero-Shot Learning From

Noisy Texts". In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 1004–1013 (cit. on p. 97).

Zhu, Yuke, Oliver Groth, Michael S. Bernstein, and Li Fei-Fei (2016). "Visual7W: Grounded Question Answering in Images". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 4995–5004 (cit. on p. 33).

Zhu, Yukun, Ryan Kiros, Richard S. Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler (2015). "Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books". In: *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pp. 19–27 (cit. on p. 20).

Zitnick, C. Lawrence and Devi Parikh (2013). "Bringing Semantics into Focus Using Visual Abstraction". In: *2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, June 23-28, 2013*, pp. 3009–3016 (cit. on p. 136).

Zitnick, C. Lawrence, Ramakrishna Vedantam, and Devi Parikh (2016). "Adopting Abstract Images for Semantic Scene Understanding". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 38.4, pp. 627–638 (cit. on p. 136).

Zoph, Barret, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le (2018). "Learning Transferable Architectures for Scalable Image Recognition". In: *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 8697–8710 (cit. on pp. 28, 46, 58).