Multimodal machine learning : complementarity of textual and visual contexts

Eloi Zablocki

École Doctorale Informatique, Télécommunications et Électronique (Paris)

DOCTORAL THESIS

Multimodal machine learning: complementarity of textual and visual contexts

Éloi Zablocki

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

Tentative defense date: October 14th, 2019

Jury composed of:

Mr. Antoine Bordes	Facebook	Examiner
Mr. Patrick Gallinari	Sorbonne Univ. — Criteo	Supervisor
Mr. Guillaume Gravier	IRISA	Reporter
Mrs. Marie-Francine Moens	KU Leuven	Reporter
Mr. Benjamin Piwowarski	Sorbonne University	Co-supervisor
Mrs. Laure Soulier	Sorbonne University	Co-supervisor
Mr. Xavier Tannier	Sorbonne University	Examiner

ABSTRACT

Research looking at the interaction between language and vision, despite a growing interest, is relatively underexplored. Beyond trivial differences between texts and images, these two modalities have non overlapping semantics. On the one hand, language can express high-level semantics about the world, but it is biased in the sense that a large portion of its content is implicit (common-sense or implicit knowledge). On the other hand, images are aggregates of lower-level information, but they can depict a more direct view of real-world statistics and can be used to ground the meaning of objects. In this thesis, we exploit connections and leverage complementarity between language and vision.

First, natural language understanding capacities can be augmented with the help of the visual modality, as language is known to be *grounded* in the visual world. In particular, representing language semantics is a long-standing problem for the natural language processing community, and to further improve traditional approaches towards that goal, leveraging visual information is crucial. We show that semantic linguistic representations can be enriched by visual information, and we especially focus on visual contexts and spatial organization of scenes. We present two models to learn grounded word or sentence semantic representations respectively, with the help of images.

Conversely, integrating language with vision brings the possibility of expanding the horizons and tasks of the vision community. Assuming that language contains visual information about objects, and that this can be captured within linguistic semantic representation, we focus on the zero-shot object recognition task, which consists in recognizing objects that have never been seen thanks to linguistic knowledge acquired about the objects beforehand. In particular, we argue that linguistic representations not only contain visual information about the visual appearance of objects but also about their typical visual surroundings and visual occurrence frequencies. We thus present a model for zero-shot recognition that leverages the visual context of an object, and its visual occurrence likelihood, in addition to the region of interest as done in traditional approaches.

Finally, we present prospective research directions to further exploit connections between language and images and to better understand the semantic gap between the two modalities.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

BoW	Bag-of-Words
BoVW	Bag of Visual Words
BLEU	BiLingual Evaluation Understudy
BPE	Byte Pair Encoding
CBOW	Continuous Bag-of-Words
CCA	Canonical Correlation Analysis
ConvNet	Convolutional Neural Network
CLEVR	Compositional Language and Elementary Visual Reasoning diagnostics
CMPlaces	Cross-Modal Places
CR	Customer Reviews
CRF	Conditional Random Field
CV	Computer Vision
DSM	Distributional Semantic Model
GAN	Generative Adversarial Network
GPUs	Graphics Processing Units
GRU	Gated Recurrent Unit
HOG	Histogram of Oriented Gradient
IR	Information Retrieval
KB	Knowledge Base
LDA	Latent Dirichlet Allocation
LSA	Latent Semantic Analysis
LSTM	Long-Short Term Memory
mNNO	mean Nearest Neighbor Overlap
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

MPQA	Multi-Perspective Question Answering
MS COCO	Microsoft Common Objects in Context
NER	Named Entity Recognition
NLP	Natural Language Processing
NMT	Neural Machine Translation
PCA	Principal Component Analysis
PMI	Point-wise Mutual Information
POS	Parts-of-Speech
QA	Question Answering
ReLU	Rectified Linear Unit
RBF	Radial Basis Function
RNN	Recurrent Neural Network
RQ	Research Question
SGD	Stochastic Gradient Descent
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

Chapter 1

INTRODUCTION

Contents

1.1 Context

Over the last three decades, there has been a large development of digital services, including data-sharing platforms, forums, on-demand streaming websites, social networks... As the cost to store a gigabyte of data has decreased from $500,000\$$ to $0.2\$$ in forty years, these industries now accumulate massive amounts of data, typically texts and images. For example, every day, it is estimated that about one billion of photos are uploaded on Facebook, and about 650 million tweets are posted on Twitter. Dealing with this huge amount of data has thus emerged as a major challenge, commonly addressed by *machine learning* approaches, which enable to extract meaningful knowledge from raw data and to interact with users.

One of the most prominent application case of machine learning is Natural Language Processing (NLP), a research field dealing with **natural language** data. NLP covers various subfields such as syntax and semantic analysis (Y. Bengio et al. 2003; Petrov et al. 2012), Information Retrieval (IR) (Salton et al. 1975), sentiment analysis (Pang et al. 2007), automatic translation (Bahdanau et al. 2015)... These problems are traditionally addressed with heuristic models, which are themselves based on simple statistics such as counting word occurrences and co-occurrences in a document (Hristea 2011). Over the last thirty years, a new idea has emerged for NLP: the *representation learning* paradigm (Deerwester et al. 1990; Y. Bengio et al. 2003; Y. Bengio et al. 2013). The key idea of representation learning is to learn a *representation* for a textual unit (a word, a sentence or a document...). Typically, these representations encode the *semantic* (*i.e.* the meaning) of the textual unit: in the space wherein the representations are, the semantic similarity between two textual unit can be measured as a spatial proximity between the representations. Several methods have then been proposed to learn representations for textual units, based on supervised (Conneau et al. 2017) or unsupervised objectives (Mikolov et al. 2013). Once a representation space is learned, it is common practice to use it for downstream NLP tasks, as it is supposed to contain rich syntactic and semantic information (Devlin et al. 2018). Learning high-quality textual representations is as a crucial challenge for the NLP community.

Another important area of the application of machine learning is to provide means of automatically understanding **images**, and related media such as videos. This is the purpose of the *computer vision* field. In particular, that aims at extracting meaningful and high-level information from abundant low-level information (pixel values). It encompasses various challenges such as detecting objects in images (localization, classification, segmentation) (Ratan et al. 1998; Borenstein et al. 2006), estimating human poses (Parameswaran et al. 2004), recognizing handwritten characters (Kae et al. 2010)... Traditionally, no (or few) learning is involved to extract images features (Lowe 2004; Dalal et al. 2005), and in a second phase, these manually designed features are fed to a machine learning algorithm which learns to perform the task of interest. In the early 2010's, huge progresses have been made in computer vision (Krizhevsky et al. 2012) thanks to three factors: (1) theoretical advances, (2) increasing computing powers, and (3) the development of large-scale public datasets. Convolutional Neural Network (ConvNet), which were invented in the 1980's (Fukushima et al. 1982; LeCun et al. 1989), have seen their usage and performance widely increased. These *deep* networks consist of successive layers which learn hierarchical visual representation of an image. For example, given the image of a person's face, the first layers can detect edges and corners, the next layers can detect larger patterns such as an eye or a nose, and the final layers can recognize face's shape (Zeiler et al. 2014). Therefore, by extracting intermediate activation values, ConvNet architectures can produce learned distributed representation for images.

Beyond the independent study of machine learning for either NLP or computer vision, the interaction of language and images is still relatively underexplored, despite a growing interest. Exploring machine learning approaches dealing with these two modalities is the focus of this thesis.

1.2 Research questions

Textual and visual modalities are trivially different by the way information is encoded: language is a discrete signal — made of words, sentences and paragraphs —, while an image is continuous and composed of spatially arranged pixels. This has the consequence that representation learning techniques for images or language are specific to the modality, and so are the produced representations which are embedded in different spaces. Based on this observation, several works have thus attempted to learn simple associations between language and images.

	Language	Images
Origin	human defined	raw signal
Atomic values	discrete (words)	continuous (pixels)
Structure	sequential	spatial
Expressed semantics	high-level	low-level
Real-world statistics	biased	accurate
Need for supervision	low	high

Table 1.1 – **Differences between linguistic and visual modalities**

This includes works that learn to combine linguistic and visual representation and works that jointly learn a multimodal representation space to embed both modalities.

Beyond the trivial difference, in the way information is encoded, we now highlight more fundamental differences between language and images (reported in Table 1.1).

- Visual data are direct depictions of the reality and are not subject to interpretation: views of objects and spatial organization of scenes in images are unequivocal, and images thus report faithful real-world statistics. However, this comes with the fact that the semantics expressed in images is only low-level. It has the consequence that learning semantic representations from images requires a lot of supervision with current approaches.

- On the other hand, language can be ambiguous, relies on context and background knowledge (*e.g.* common-sense), and consequently is not an unbiased transcription of the reality as, for example, humans tend not to mention unsurprising facts. The latter is known as the *reporting bias* (Gordon et al. 2013). However, unlike images, language can refer to high-level concepts. Besides, there exist several approaches to learn linguistic semantic representation without supervision.

These fundamental differences between language and images hint at complementarity of both modalities. This is the core problem of the thesis that we decouple into two complementary research questions presented below and illustrated in Figure 1.1.

The first axis which we explore in the thesis addresses the following research question:

Can language be grounded in the visual world?

The issue of the *reporting bias*, *i.e.* the fact that language contains biased real-world statistics and lacks common-sense, can be alleviated by leveraging information from other resources, typically images which do not suffer from this bias. This first axis is illustrated in green in Figure 1.1. Several approaches have explored this

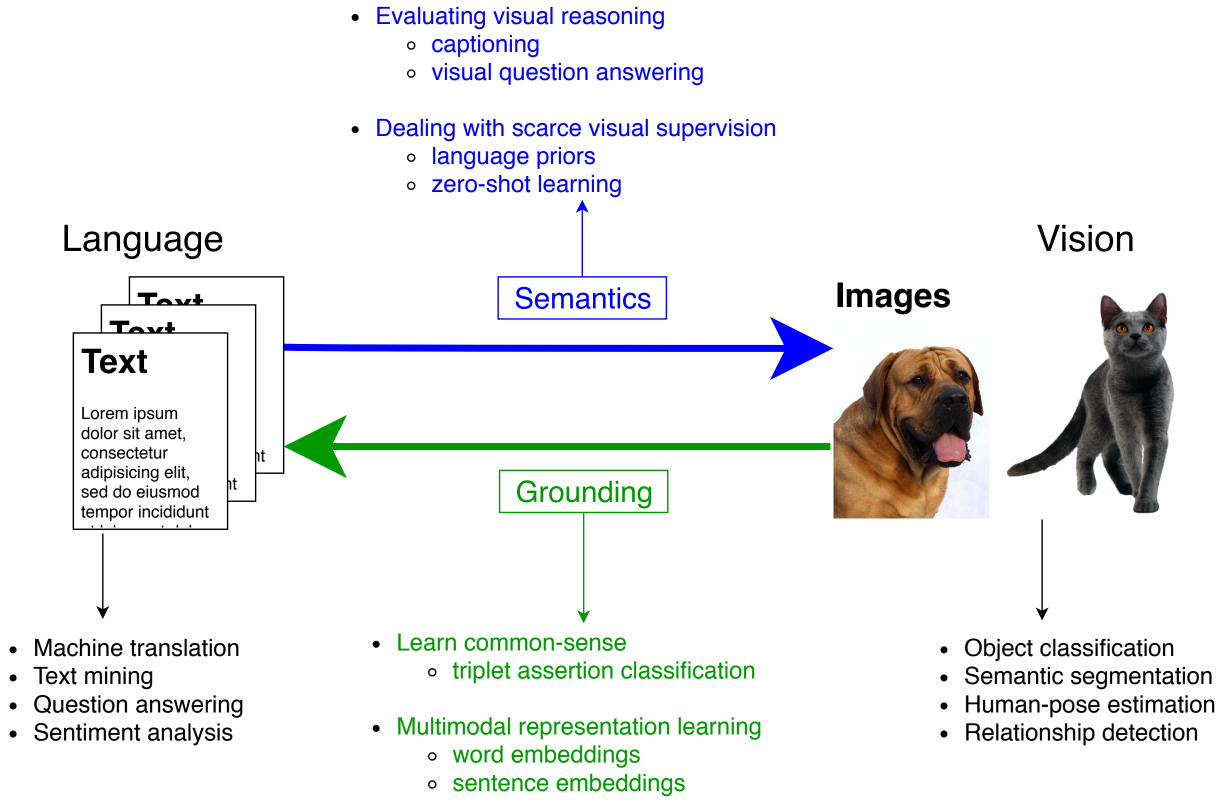


Figure 1.1 – **Overview of multimodal machine learning with language and images**. In this thesis, we tackle two complementary research questions. (1) *Can language be grounded in the visual world?* (illustrated in green), we give elements of response in Chapter 3 and Chapter 4.
(2) *Can language help visual recognition?* (illustrated in blue), question tackled in Chapter 5.

line of research and they usually focus on incorporating visual information into distributed linguistic representations, *i.e.* learning multimodal general-purpose representations. Using images — which typically provide common-sense knowledge — allows us to enrich semantic representations of objects, for example by giving information about color, shape, or typical surroundings of these objects. In this thesis, some contributions are directed towards that goal, and we propose two models to learn grounded representations for words (Chapter 3) and sentences (Chapter 4).

The second axis takes the opposite perspective and addresses the following research question:

Can language help computer vision?

Based on the fundamental differences between language and images, we distinguish two orthogonal approaches where language can be leveraged to benefit the visual modality (illustrated in blue in Figure 1.1):

- Language can play a role to augment the visual understanding capacities of a model. Indeed, assuming that linguistic semantic representations contain visual information, these semantic representations can help recognizing objects or reasoning with visual situations. This is of particular interest given that traditional visual recognition systems rely on a lot of supervised data, while it is possible to learn semantic representation with unsupervised machine learning approaches on texts. Typically, when visual supervision is scarce, leveraging linguistic representations shows great benefits (Frome et al. 2013; R. Yu et al. 2017). In the extreme case, when visual supervision is lacking and that certain objects are not seen at all during training, it is possible to use semantic representations to recognize the unseen objects. This scenario corresponds to the *zero-shot* object recognition, which we tackle in Chapter 5.

- Language can serve as a way to evaluate visual models. Models that can express in natural language the content of an image can demonstrate their capacity to extract high-level semantics from images, and their ability to reason with visual content. This is the fundamental hypothesis that motivates the need to evaluate visual systems with natural language, and it has lead to the development of the image *captioning*, and the Visual Question Answering (VQA) task. While it is discussed in Chapter 2, this is not the focus of the thesis.

1.3 Contributions and outline of the thesis

The contributions of the thesis are outlined as follow:

- In Chapter 2, we present background multimodal machine learning approaches for text and images. We first review unimodal machine learning methods, either in the case of text or images, and then present motivations for leveraging textual and visual modalities altogether. We detail a first line of works which attempts to incorporate visual semantics to NLP, for example to ground the meaning of words or sentences, or to learn common-sense. Conversely, we present the opposite approach, where the goal is to use natural language to either help visual understanding and reasoning, or to cope with the fact that most of visual learning systems rely on a strong supervision signal.

The outline of the rest of the thesis follows these two lines of research: in Chapter 3 and Chapter 4 we focus on incorporating visual semantics into linguistic representations (first axis, illustrated in green in Figure 1.1), and in Chapter 5, we rather use linguistic representations to augment visual recognition capacities (second axis, illustrated in blue in Figure 1.1):

- In Chapter 3, we build upon works which incorporate visual semantics to *word* representations. While previous works usually consider the visual appearance of objects to enhance representations, we investigate whether other visual sources of information contain complementary semantics. We show that the visual surroundings (*i.e.* the visual context) of objects in an image, as well as the spatial organization of a scene, can further improve word representation. In practice, we build upon the `skip-gram` algorithm, which traditionally considers *textual* contexts to learn the meaning of a word, and extend it to *visual* and *spatial* contexts. This work has been published: Eloi Zablocki, Benjamin Piwowarski, Laure Soulier, and Patrick Gallinari (2018). "Learning Multi-Modal Word Representation Grounded in Visual Context". In: *AAAI 2018*.

- In Chapter 4, the goal is to incorporate visual semantics into *sentence* representations. Some linguistic and their matching visual phenomena only take place at a sentence level, as sentences can be visually ambiguous, carry non-visual information, or have a wide variety of paraphrases and related sentences describing a same scene. We thus propose to transfer visual information to the sentence embeddings through an intermediate space in which we define two complementary objectives to explicitly and implicitly incorporate visual semantics to the learned representations. This work corresponds 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*.

- In Chapter 5, we present a work aiming at augmenting visual recognition systems thanks to linguistic priors, in contrast with Chapter 3 and Chapter 4 which focus on enriching textual representations with visual semantics. We consider the *zero-shot* scenario for object recognition, where some classes are not seen during training. This problem is traditionally addressed by leveraging auxiliary linguistic information, *e.g.* in the form of class label embeddings, which are supposed to contained visual information about the direct appearance of objects. We show that these textual representations contain additional visual information, about the typical visual surroundings of objects and the likelihood of visual occurrence of objects. We explicitly use this information to tackle the *context-aware* zero-shot recognition scenario that we propose. 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 6, we conclude the thesis, by summarizing the different contributions and by presenting mid-term and long-term future research directions.

Chapter 2

MULTIMODAL MACHINE LEARNING WITH IMAGES AND TEXT: BACKGROUND

Contents

Chapter abstract

In this chapter, we present machine learning approaches for dealing with language and visual data, either separately or jointly. We give an overview of issues and questions that arise when dealing with language and vision, and review multimodal machine learning challenges to tackle these issues. We start by presenting techniques to represent language (Section 2.1.1) and visual data (Section 2.1.2). We then turn to classical multimodal learning challenges: learning associations between language and images (Section 2.1.3). We finally review machine learning approaches that exploit connections and leverage complementarity between language and vision, and thus consider the two following view points which supplement each other:

- *Natural language understanding capacities can be augmented with the help of the visual modality, as language is naturally grounded in the visual world (Section 2.2).*

- *Conversely, integrating language with vision brings the possibility of expanding the horizons and tasks of the computer vision literature (Section 2.3), either as a mean to evaluate computer vision systems (Section 2.3.1) or when supervision is a limiting factor (Section 2.3.2).*

Disclaimer: *The purpose of this chapter is to give an overview of machine-learning research questions and approaches dealing with texts and images. Given the number of works that fall into this scope, we cannot give in-depth explanations for every technique mentioned in this chapter, and we rather focus on (1) presenting the major issues and approaches and (2) detailing approaches that are directly linked to the remaining chapters*

2.1 Machine learning with language and vision

Traditional machine learning pipelines rely on preprocessed inputs, which are computed with manual techniques known as *feature engineering*. In contrast, the *deep learning* paradigm has largely contributed to the rise of *representation learning*, by leveraging large amounts of data. *Representation learning* designates techniques that enable the automatic discovery of semantic representations, *i.e.* embeddings in a vector space such as $\mathbb{R}^d$, where meaningful features are encoded.

Across many different tasks and domains, learning data representation has shown to be more efficient (in downstream model performances) and effective (easier and cheaper to produce) than having experts designing handcrafted features (Y. Bengio et al. 2013; LeCun et al. 2015). The representation $f_\theta(x)$ of a data input x is typically obtained by finding the optimal θ that minimizes a loss function (*e.g.* the negative log-likelihood of a dataset). When the objective function is differentiable with respect θ, the optimal value for θ is usually found with a gradient descent algorithm (LeCun et al. 1989).

In this thesis, we focus on representation learning techniques dealing with language (*e.g.* text) and vision (*e.g.* images). First, we present methods to learn monomodal representations for both linguistic data (Section 2.1.1) and visual data (Section 2.1.2). We then present how independent monomodal representations can further be mingled (Section 2.1.3), either by fusing them (Section 2.1.3.1), or by building a multimodal shared space (Section 2.1.3.2).

2.1.1 Representations for Natural Language Processing (NLP)

Textual data consists of compositional sequences of symbols (characters). Characters can be grouped to form words, words can form sentences, which can then be aggregated into paragraphs etc. . . . Natural language obeys syntactic and grammatical rules, and one of the main challenge for natural language understanding is to extract semantics from words and their relations.

In this section, we review traditional and more recent approaches that deal with language data. This includes the Bag-of-Words (BoW) model, language models, and representation learning techniques for words, sentences and documents. Moreover, we regroup in Table 2.1 a list of NLP tasks and applications that are impacted by the quality of the input representation.

Task	Description	References
Sentiment analysis	Identify and study subjective information such as affect, sentiment or emotional state	(Pang et al. 2007; Maas et al. 2011; Pontiki et al. 2016)
Question Answering (QA)	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 one	(Bahdanau et al. 2015; Artetxe et al. 2018; Lample et al. 2018a)
Dialog systems	Converse with a human with a coherent structure	(Vinyals et al. 2015; Sordoni et al. 2015)
Parts-of-Speech (POS) tagging	Label a word in a text as corresponding to a particular part of speech (*e.g.* noun, verb, adjective . . .)	(Petrov et al. 2012; Nguyen et al. 2016)
Named Entity Recognition (NER)	Find and classify named entities in text into pre-defined categories, such as persons, organizations, locations, . . .	(Lample et al. 2016; Moreno et al. 2017)
Text summarization	Shorten a text into a summary containing the major points of the original document	(Rush et al. 2015; Aries et al. 2019)
Speech recognition*	Recognize and translate spoken language into text	(Katz 1987; Vukotic et al. 2015; Deena et al. 2019)
Handwriting recognition*	Recognize and translate handwritings into text	(Frinken et al. 2012)

Table 2.1 – **Examples of NLP tasks and applications** which are conditioned by successfully capturing the meaning and semantics of words and documents. *: Even though these tasks do not handle texts, the design of modern handwriting/speech recognition systems heavily relies on a good language model.

2.1.1.1 One-hot encoding and BoW model

The simplest way to encode a word w into a vector $t_w \in \mathbb{R}^d$ is known as *one-hot encoding*. Considering a dictionary of N words, and assuming that the word w is the i-th word of this dictionary, the *one-hot encoding* of w is a binary, sparse word vector $t_w \in \{0, 1\}^N$, where all coordinates are 0 except for the i-th coordinate which equals 1.

The representation of a sentence or a document can be obtained with the BoW model. This model simply computes the sum or average of one-hot encodings of each word contained in the document (Buscaldi et al. 2006; Metzler 2008).

The BoW model has the advantage of being simple, as no learning is involved to build the representation, but it comes with several limitations, some of them being inherently related to *one-hot encoding*:

- The word order is lost in the final encoding — this can be problematic as sentences containing the same words in different orders (hence having possibly different meanings) will be encoded the same: "Alice likes Bob" vs "Bob likes Alice". To avoid this problem, a possibility is to take the word order into account for example in n-gram or language models (detailed below in Section 2.1.1.2).

- Representations can get very big as they grow with the dictionary size N, which can be several millions. In addition, it is not possible to add a new word *a posteriori*. A way to circumvent this problem is to embed all words into a space with a fixed and predefined size, this is the idea used to learn word representation (detailed in Section 2.1.1.3).

- Representations do not encode any notion of semantic and syntax similarity, as the euclidean distance between any pair of words is constant: *a priori* the word "dog" is as similar to a "truck" as it is to a "cat". To bypass this issue, it is either possible to use relational semantics encoded in external structured resources such as Knowledge Base (KB) (Speer et al. 2017), or to learn word vectors, as detailed in Section 2.1.1.3.

2.1.1.2 Language models

A *language model* is a probability distribution over sequences of words (or characters). It computes a probability $P(w_1, w_2, \cdots, w_n)$ for any sequence of n words. To design good language models, the main issue that needs to be addressed is the data sparsity, as most of possible word sequences are never observed.

n-gram language model A first possibility to design a language model is to use n-grams (ordered sequence of n words) statistics. Based on the Markovian assumption that closer words in a word sequence are statistically more dependent (Y. Bengio et al. 2003), a n-gram language model computes the probability of having word w_i occurring in a text, when only the $n-1$ previous words are given: $P(w_i \mid w_{i-1}, \cdots, w_{i-(n-1)})$. This corresponds to a probabilistic Markov model which models sequences given n-gram statistics computed over a whole textual corpus (count-based technique) (Shannon 1951). This modeling has the advantage of being simple to compute and to consider the word order. However, this simple model has several limitations. First, the number of n-grams grows with N^n which is problematic given limited memory resources. Instead, recent approaches project words to a continuous space (see Section 2.1.1.3) with both feedforward and recurrent neural networks (see below). Second, like in the case of the BoW model, no notion of syntactic and semantic similarity is learned. Third, if a n-gram is not observed in the training set, zero-probability will be given to it, which will cause a zero probability over entire sequences. As a workaround, smoothing techniques assign non null probabilities to n-grams unseen in the training corpus (Jeffreys 1948; Nadas 1984; Manning et al. 2008).

Neural language model Neural language models learn continuous representations of words to make their predictions. For the probability P that we wish to model on the word sequence, the chain rule gives:

$$P(w_1, w_2, \cdots, w_n) = \prod_{i=1}^{n} P(w_i \mid w_1, \dots, w_{i-1}) \tag{2.1}$$

Recurrent Neural Network (RNN) can be used to summarize the history of previous words, within a fixed-size state h_i (Y. Bengio et al. 2003), and make the following hypothesis:

$$P(w_{i+1} \mid w_1, \dots, w_{i-1}) \propto g(w_{i+1}, h_{i-1}) \tag{2.2}$$

where g is typically expressed as:

$$\begin{cases} g(w_{i+1}, h_{i-1}) = \exp(W_o h_i + b_o) \\ h_i = \phi(w_{i+1}, h_{i-1}) \end{cases} \tag{2.3}$$

where W_o and b_o are trainable weights and ϕ is a recurrent function.

The properties of the RNN highly depend on the choice of the recurrent function ϕ. For example, simple recurrent networks use the following function: $\phi(w, h) = \sigma(W_i w + W_r h + b)$, where W_i, W_r and b are trainable weights and σ is the sigmoid or tanh function (Elman 1990). More sophisticated recurrent functions have also been proposed, such as the Long-Short Term Memory (LSTM) (Hochreiter et al. 1997) which alleviates the problem of vanishing and exploding gradient (Y. Bengio et al. 1994) or the Gated Recurrent Unit (GRU) (Cho et al. 2014) which has fewer parameters and shows similar performances. Besides RNN, Convolutional

Neural Network (ConvNet) can also be used for language modeling (Schwenk et al. 2017); they can benefit from the depth of the architecture to learn hierarchical representation for language: similarly to their use in image, such architectures are very useful for classification tasks. Over *n*-gram models, neural language models have the advantage to be able to deal with long range dependencies (Schäfer et al. 2006), and to avoid the use of smoothing techniques thanks to the use of distributed representations as inputs (see Section 2.1.1.3).

Latest advances for language modeling include the use of sub-word tokens to avoid out-of-vocabulary problems when a word has not be seen during training, as for example the Byte Pair Encoding (BPE) proposed by Sennrich et al. 2016. Moreover, more and more sophisticated neural architectures are used. For example, in machine translation, attention mechanisms are widely used (Bahdanau et al. 2015) ; in language modeling, pre-trained transferable language models are now built with a transformer architecture (Vaswani et al. 2017): `BERT` (Devlin et al. 2018) and `GPT-2` (Radford et al. 2019) are famous recent examples.

2.1.1.3 Word vectors

Part of the success of neural language models is that they are based on the hypothesis that each word can be represented by a vector, allowing them to significantly outperform *n*-gram models (Schwenk 2007; Mikolov et al. 2011). Geoffrey E Hinton 1986 originally proposed to learn *distributed representations* for symbolic data. This idea was applied for statistical language modeling (Schütze 1992; Y. Bengio et al. 2003; Y. Bengio 2008), where a representation[1] $t_w \in \mathbb{R}^d$ is learned for each word w (d denotes the dimension of the vector space, typically d ranges from 100 to 1000).

A Distributional Semantic Model (DSM) represents the *semantic* (*i.e.* meaning) of words with vectors encoding the distribution of the word contexts in the corpus (Baroni 2016). A DSM leverages large text corpora under the *Distributional Hypothesis* (Harris 1954), *i.e.* the assumption that *words that occur in similar contexts should have similar meanings*. They produce fixed-length vectorial representation for words based on their co-occurrences in text corpora. In practice, a DSM computes the representation of words through an implicit or explicit factorization of a co-occurrence matrix (Levy et al. 2014c). Well-known DSM are `GloVe` (Pennington et al. 2014) and `Word2Vec` (Mikolov et al. 2013). They differ in that `Word2Vec` is a "predictive" model, while `GloVe` takes a "count-based" approach (Baroni et al. 2014). In `Word2Vec` , words are either predicted given their context (Continuous Bag-of-Words (CBOW) model) or vice-versa (`skip-gram` model).

We now present the `skip-gram` model in more details, as it is the basis of one of our models. Given a word e in a text corpus $\mathcal{D}$, and a word c that occurs in a window of size l centered on e, we say that e is an *entity* and c is its *context*. We

[1] Throughout this thesis, the words *representation* and *embedding* will be used interchangeably.

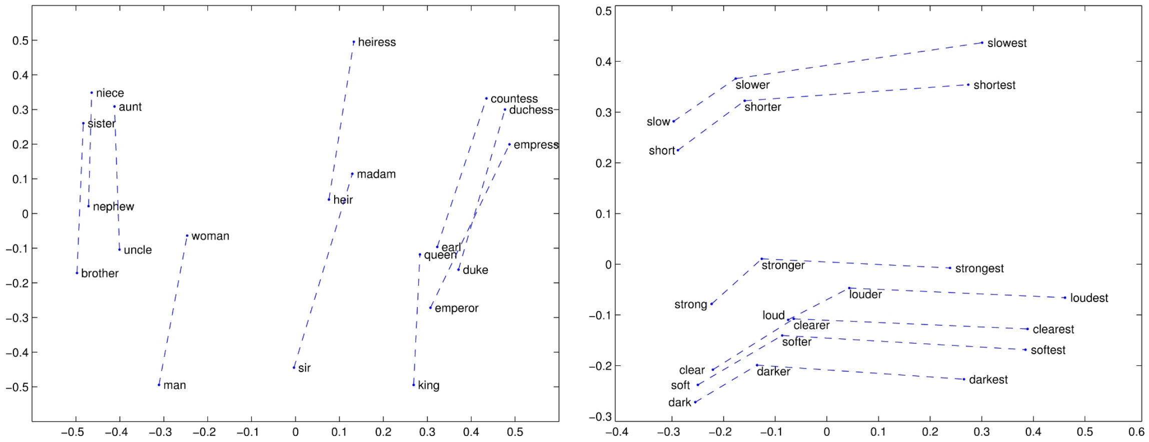


Figure 2.1 – **Arithmetic relations observed in a word representation space**. `GloVe` representations are projected with Principal Component Analysis (PCA) to a 2-dimensional space. Relation between word pairs (such as man/woman or comparative/superlative) are encoded by vectors having consistently the same direction and orientation. Illustration taken from (Pennington et al. 2014)

note $\mathcal{C}_e$ the set of context words for the entity e and D the binary random variable which equals 1 if e and c are a positive pair (c is a context of e) and 0 if e and c are a negative pair (c is a word sampled randomly over the vocabulary). The aim is to learn word representations that maximize the probabilities $P(D = 1 \mid e, c; \theta)$ for positive (e, c) pairs (*i.e.* $D = 1$) et $P(D = 0 \mid e, c; \theta)$ for negative (e, c) pairs (*i.e.* $D = 0$). θ denotes the trainable weights, *i.e.* word embeddings in our case. The `skip-gram` algorithm consists in finding the parameters θ that maximize the following cross-entropy:

$$\arg\max_{\theta} \sum_{e \in \mathcal{D}} \sum_{c \in \mathcal{D}} \Big[D \log(P(D = 1 \mid e, c; \theta)) + (1 - D) \log(P(D = 0 \mid e, c; \theta)) \Big] \tag{2.4}$$

In practice, the original `skip-gram` algorithm models the probability P as follow: $P(D = 1 \mid e, c, ; \theta) = \sigma(u_c^\top t_e)$, where σ is the sigmoid function: $\sigma(x) = \frac{1}{1+e^{-x}}$. In this modeling, two representations, t_e and u_c, are used for each word, according to their role, entity e or context c respectively.

With some approximations, explained in the original paper (Mikolov et al. 2013), the final objective is finally obtained:

$$\arg\max_{\theta} \sum_{e \in \mathcal{D}} \sum_{c \in \mathcal{C}_e} \Big[\log \sigma(u_c^\top t_e) + \sum_{c^-} \log \sigma(-u_{c^-}^\top t_e) \Big] \tag{2.5}$$

The negative contexts c^- are randomly sampled according the word frequency. This optimization technique is named the *negative sampling*. The trainable parameters θ are two matrices $T \in \mathbb{R}^{N \times d}$ and $U \in \mathbb{R}^{N \times d}$ which contain the representation of words, either as entities, or as context (the i-th row of T and U are the respective representations t_w and u_w of the i-th word w of the dictionary). More details can be found in (Goldberg et al. 2014). At the end of training, the matrix T containing the word representation is used to provide embedding for each word and the matrix U is discarded.

Levy et al. 2014c show that `GloVe` and `Word2Vec` are variants of PCA, Singular Value Decomposition (SVD) and Latent Semantic Analysis (LSA) algorithms, in the sense that all of these methods factorize a term co-occurrence matrix. Moreover, Levy et al. 2015 observe that well-tuned SVD factorization can outperform `skip-gram` and `GloVe` algorithms which reveals that much of the performance gains of word embeddings are due to hyperparameter optimizations rather than the embedding algorithms themselves.

Empirically, one can make interesting observations on the learned semantic word vector space. First, semantically similar words are located in the same regions of the representation space (*e.g.* dog breeds form a cluster). Second, some relations between words are linear: this is the case for example the *man:woman* and the *adjective:comparative:superlative* relations (see illustration in Figure 2.1). Third, for a given word representation algorithm, the word spaces for two distinct languages show similar structures ; this enables the possibility to perform semi-supervised or fully unsupervised bilingual lexicon induction (Smith et al. 2017; Lample et al. 2018b).

Finally, several improvements have been proposed to learn better word representation, such as (1) using Gaussian embeddings to account for the variance of the meaning of words (Vilnis et al. 2015), (2) using extra information provided by KB (Tian et al. 2016), or (3) learning *contextualized* word representation (Peters et al. 2018; Devlin et al. 2018), where the representation of a word depends on its context. The particular case of using images to learn grounded word representations is detailed in Section 2.2.2.1.

2.1.1.4 Sentence vectors

Several approaches have been proposed to learn distributed semantic representation for sentences (Hill et al. 2016; Kiros et al. 2015). This is more challenging than learning word representations, since sentences are inherently different than words due to their sequential and compositional nature. 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. 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. 2015) or question answering (Sagara et al. 2014).

On the one hand, supervised techniques produce task-specific sentence embeddings. For example, in a classification context, they are built using recurrent

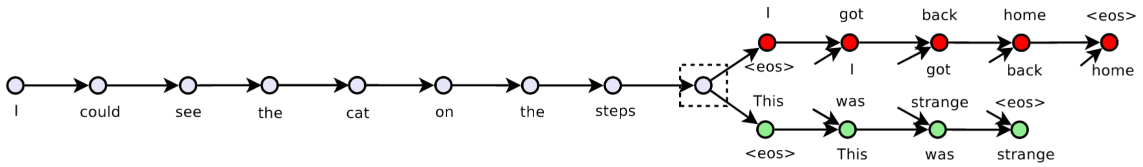


Figure 2.2 – `SkipThought` **architecture**. Image taken from (Kiros et al. 2015)

networks with LSTM (Hochreiter et al. 1997), recursive networks (Socher et al. 2013), ConvNet (Kalchbrenner et al. 2014), or self-attentive networks (Z. Lin et al. 2017; Conneau et al. 2017).

On the other hand, unsupervised methods aim at producing more general, universal, and task-independent sentence representations, given large text corpora. Examples of unsupervised techniques include models such as `FastSent` (Hill et al. 2016), `QuickThought` (Logeswaran et al. 2018), Word Information Series (Arroyo-Fernández et al. 2019), Universal Sentence Encoder (D. Cer et al. 2018), or `SkipThought` (Kiros et al. 2015).

For instance, `QuickThought` (Logeswaran et al. 2018), `SkipThought` (Kiros et al. 2015) and `FastSent` (Hill et al. 2016) are based on the distributional hypothesis (Harris 1954) applied to sentences, i.e. *sentences that appear in similar contexts should have similar meanings*. Given three consecutive sentences S_{i-1}, S_i, S_{i+1}, in the `FastSent` model, the representation s_i of the sentence S_i is the sum of its word embeddings $s_i = \sum_{w \in S_i} t_w$. The learning objective is similar to the `skip-gram` objective (Equation 2.5): predict the words of the adjacent sentences using a negative sampling loss:

$$\arg\max_{\theta} \sum_i \sum_{w \in S_{i-1} \cup S_{i+1}} \left[\log \sigma(u_w^\top s_i) + \sum_{w^-} \log \sigma(-u_{c^-}^\top s_i) \right] \tag{2.6}$$

In the `SkipThought` model, a sentence is encoded with a GRU network (Cho et al. 2014), and two GRU decoders are trained to reconstruct the adjacent sentences in a dataset of ordered sentences. The `SkipThought` architecture is illustrated in Figure 2.2. The `QuickThought` model is a related architecture, but instead of learning to reconstruct surrounding sentences, a classification objective is optimized to distinguish context sentences from other (negative) sentences.

Document-vectors The next level of representation is at the *document* level. Historically, the Term Frequency-Inverse Document Frequency (TF-IDF) method is used in Information Retrieval (IR) and in data mining to encode the meaning of a document (Salton et al. 1984; Jones 2004; Vulic et al. 2015). In a TF-IDF model, the importance of a word to a document, with respect to a corpus, is computed as follows: the importance of a word in a document increases linearly with the number of occurrence of the word in the document, but it is rescaled by the frequency of the word in the global collection, such that frequent words are not considered as 'important'. This method is easy to compute and provides a simple way to compute the similarity between two documents. However, no word order is considered, nor does the model encode semantics and co-occurrence information.

More recently, neural approaches have been proposed, as with the `Doc2Vec` model (Le et al. 2014) which can generate unsupervised paragraph or documents representations. In this model, a target word is predicted given (1) neighboring words (like in the `skip-gram` model) and (2) a unique document vector learned for each document. This model gives competitive results with TF-IDF representations, and Latent Dirichlet Allocation (LDA) modeling (Blei et al. 2001), on a variety of text understanding tasks (Dai et al. 2015). This model has been extended for example to be more memory and time-efficient (M. Chen 2017).

2.1.2 Representations for computer vision

2.1.2.1 Before the ConvNet era

In traditional computer vision pipelines, no (or few) learning is involved to compute image representations. The broad idea is to compute local descriptors, which are then aggregated to form global image features. For example, these local descriptors could be obtained by convolving kernels on images, with kernels coming from filter banks. Based on convolving gaussian kernels on images at different scales, the Scale-Invariant Feature Transform (SIFT) (Lowe 2004) descriptors provide features invariant to rotations and image scaling. The obtained features have been shown to be robust across shifts in intensity and illumination, change in 3D viewpoint and distortions to some extent. As another example of feature descriptor, the Histogram of Oriented Gradient (HOG) (Dalal et al. 2005) counts gradient orientations in small image regions, and the obtained features are refined for edge detections and object classification. HOG descriptors provide reliable features for machine learning tasks such as the human detection task as in the original paper (Dalal et al. 2005). Local image descriptors can be aggregated in Bag of Visual Words (BoVW) models (Qiu 2002). The BoVW model for image representation is analog to the BoW model for representing text, but in the case of images, four steps are involved in the process (illustrated in Figure 2.3):

1. feature detection (detect image keypoints),
2. feature description (usually with SIFT or HOG),
3. codebook generation (cluster feature descriptors to learn codewords),
4. represent an image as a BoW with codewords.

Once an image is represented, any discriminative model such as a naive-Bayes model or a Support Vector Machine (SVM) are then used to optimize performances on a task of interest.

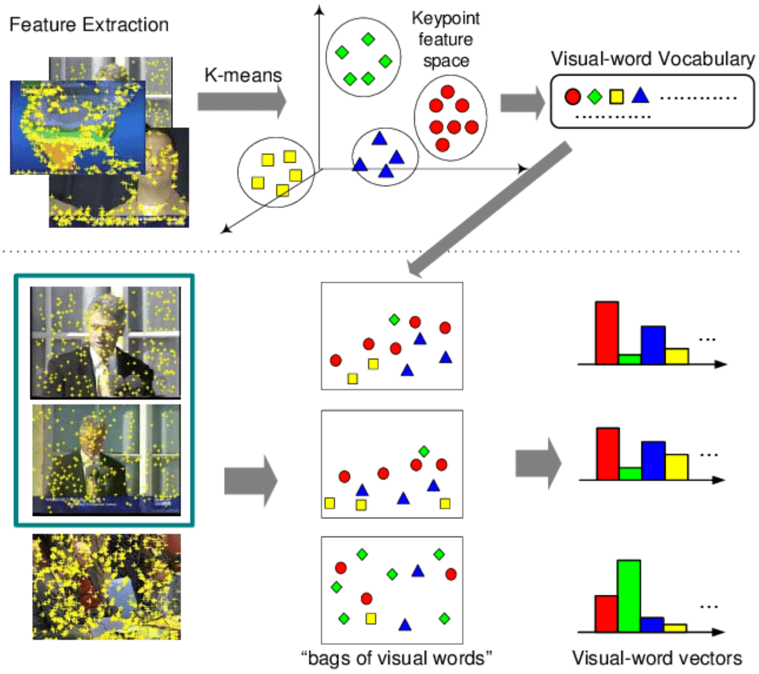


Figure 2.3 – **Illustration of the BoVW approach.** Image taken from (Jiang et al. 2010)

This traditional pipeline has several limitations:

- like for the BoW model that represents documents, the BoVW model ignores the spatial structure of image patches.
- Designing good feature descriptors is task-dependent, and, generally expert knowledge is incorporated in the used filter banks (Y. M. Lu et al. 2007). This creates models that poorly generalize to new domains and that are costly to design as intensive human intervention is needed.

2.1.2.2 Image representation with ConvNet

As mentioned in the introduction, the deep learning revolution comes from the ability of neural networks to learn rich and expressive data representation. This emerging paradigm has shown to be much more efficient and effective than directly using handcrafted features. The most popular neural networks that learn image representations are called ConvNet, as they convolve learned image patches on the images to extract hierarchical deep representations. For example, given the image of a person's face, the first layers can detect edges and corners, the next layers can detect larger patterns such as an eye or a nose, and the final layers can recognize face's shape (Zeiler et al. 2014).

Inspired by biological processes, and how certain neurons in the visual cortex were found to fire when certain images were shown, Fukushima et al. 1982

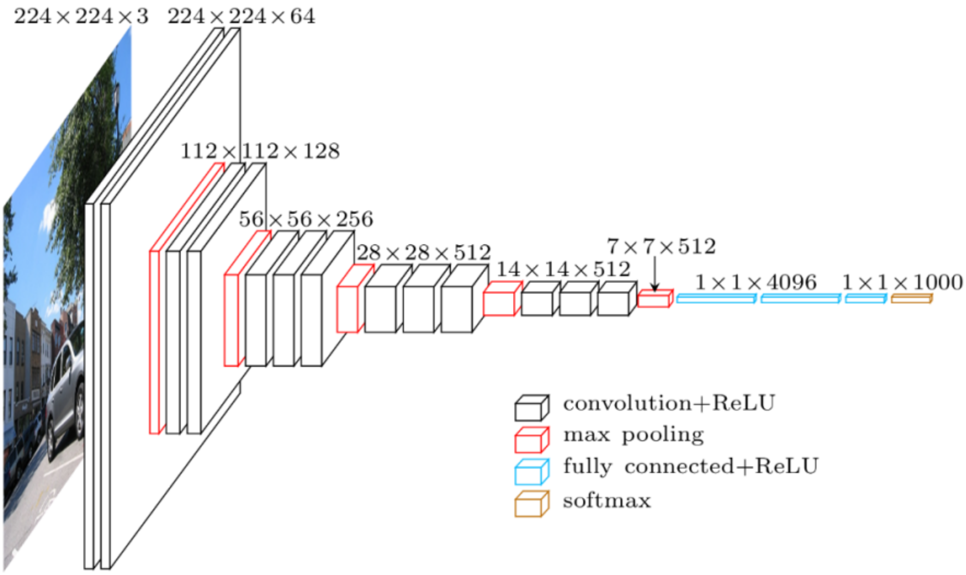


Figure 2.4 – **Illustration of a ConvNet.** In this case a VGG network (Simonyan et al. 2015), image taken from (Durand 2017).

proposed the seminal idea about ConvNet. A ConvNet architecture is based on a stacking of three types of layers, as illustrated in Figure 2.4:

- convolutional layer, where shared and learned patches are convolved over the image,
- pooling layer, a non-linear down-sampling layer that reduces the number of input data and allow for more shift invariance,
- non-linearities, usually a Rectified Linear Unit (ReLU) (Glorot et al. 2011), where $\text{ReLU}(x) = \max(0, x)$.

Despite being introduced in the 80s, ConvNet have become widespread only from the 2010s. On a theoretical point of view, the rise can be explained by the development of the *backpropagation* algorithm (LeCun et al. 1989), regularization techniques such as *dropout* (Srivastava et al. 2014) and the use of suitable activation functions such the ReLU (Glorot et al. 2011). In addition, practical reasons also explain the rise of ConvNet-based methods, such as the use of large datasets (*e.g.* Microsoft Common Objects in Context (MS COCO) (T. Lin et al. 2014), *ImageNet* (Deng et al. 2009), and *Visual Genome* (Krishna et al. 2017)), as well as fast implementations of convolutions and linear algebra computations on Graphics Processing Units (GPUs).

In 2012, the *ImageNet* challenge was won by a ConvNet-based method which outperformed previous approaches by a large margin (Krizhevsky et al. 2012). Since then, ConvNet-based architectures have won all of the subsequent *ImageNet* competitions. The convolutional architecture have been improved, with deeper and larger layers, such as with the `VGG` network (Simonyan et al. 2015), the `Inception` network (Szegedy et al. 2016) and `ResNet` network (K. He et al. 2016).

Not only does the use of ConvNet has spread to a wide variety of tasks in computer vision — object recognition, detection and segmentation, image generation with Generative Adversarial Network (GAN), and video analysis —, but ConvNet have also been widely adopted in other fields such as signal processing, NLP and games (*e.g.* the game of *go*).

2.1.3 From monomodal to multimodal representations

Textual and visual modalities are trivially different by the way information is encoded: language is made of words, sentences and paragraphs, while an image is composed of spatially arranged pixels with continuous values. This has the consequence that the representation learning techniques presented in Section 2.1.1 and Section 2.1.2 are modality-specific, and so are the produced representations. However, a natural goal is to learn associations between language and images. More specifically, given modality-specific representations for each of the textual and visual modalities, the two following questions emerge:

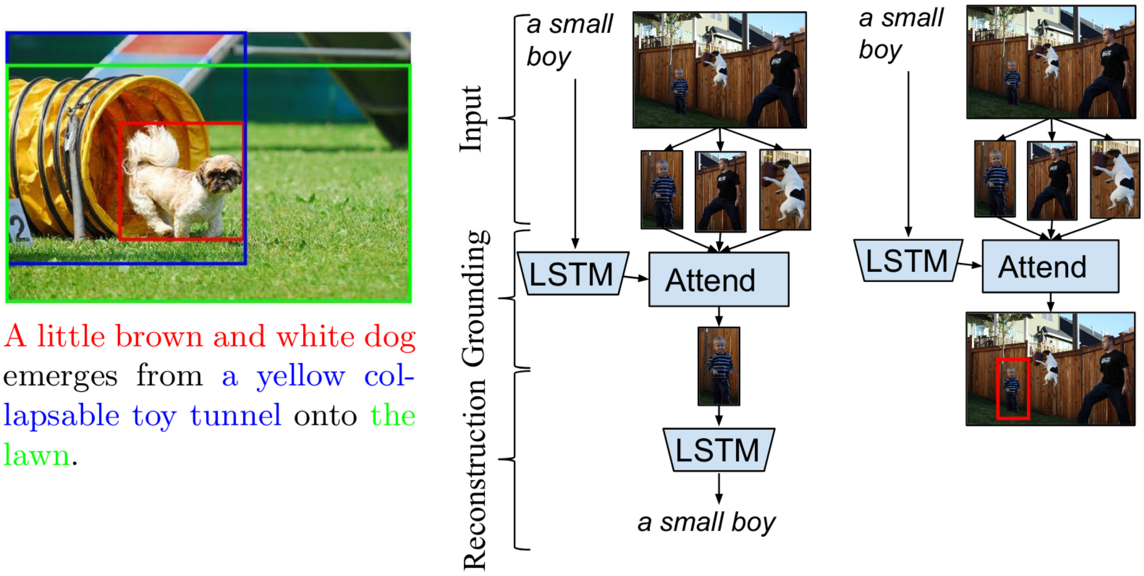
- How to combine/merge linguistic and visual representations? The *multimodal fusion* of textual and visual representations is discussed in Section 2.1.3.1.

- How to learn a shared, modality-independent, space to represent data from both modalities? Accordingly, we discuss the construction of a *multimodal shared space*, where both textual and visual representations are embedded, in Section 2.1.3.2.

These techniques can be seen as the first step towards grounded language learning methods that are presented in Section 2.2. The latter aim at learning better text representation given multimodal data, while this section presents works that focus on the bridging both visual and textual spaces so as to perform multimodal tasks like retrieval, question answering, etc. . .

2.1.3.1 Multimodal fusion

The aim of the *multimodal fusion* is to merge two mono-modal representations into a multimodal embedding. More precisely in our context, the fusion designates an operation function which takes a textual and a visual representation as inputs and outputs a multimodal representation. For example, in Visual Question Answering (VQA) the linguistic vector can be the embedding of a question and the visual vector can be the representation of an image. Multimodal fusion has several applications, as in the *visual grounding* task, which consists in localizing


A little brown and white dog emerges from a yellow collapsable toy tunnel onto the lawn.
Input
a small boy
Grounding
LSTM
Attend
Reconstruction
LSTM
a small boy
a small boy
LSTM
Attend

(a) Predicted grounding. (b) Training time. (c) Test time.

Figure 2.5 – **The *visual grounding* task**. The goal is to localize a textual phrase within an image. Image taken from (Rohrbach et al. 2016).

free-form textual phrases within an image. Rohrbach et al. 2016 tackle the visual grounding task with the *GroundeR* model, which computes compatibility of an image region with a phrase with multimodal fusion. Their model, illustrated in Figure 2.5, can be used with various levels of supervision in the annotations provided (unsupervised, semi-supervised, fully-supervised). Beyond this example, we will review special cases of multimodal fusion in Section 2.2.2 for learning word and sentence representations and applications such as VQA in Section 2.3.1.2.

Formally, given two input vectors $(t, v) \in \mathbb{R}^{d_t \times d_v}$, the multimodal fusion involves a function f_θ, parametrized by trainable weights θ, to merge t and v into a multimodal vector $m \in \mathbb{R}^{d_m}$. The simplest fusion functions are concatenation: $f_\theta(t, v) = t \oplus v$, where $\oplus$ designates the concatenation operator and the element-wise sum (resp. multiplication) $f_\theta(t, v) = t \odot v$ which requires that $d_t = d_v$, and where $\odot$ designates the element-wise sum (resp. multiplication). In any case, the produced vectors can be fed to a Multi-Layer Perceptron (MLP), which contains the trainable weights θ, to obtain the final learned representation $m \in \mathbb{R}^{d_m}$.

To allow for more complex interactions to happen when merging embeddings from the textual and visual modalities, *bilinear models* have been proposed where a tensor $T \in \mathbb{R}^{d_t \times d_v \times d_m}$ is used. However, learning a full tensor T becomes intractable as the number of parameters is quadratic in the input dimensions (Ben-younes et al. 2019). Under the umbrella of the VQA task, a recent line of works thus proposes to learn a tractable tensor T, for example by using Tucker decomposition techniques (Ben-younes et al. 2017) or a stack of low-rank matrices (Z. Yu et al. 2017).

2.1.3.2 Multimodal shared space

Embedding language and image in a *shared* semantic space is a fundamental goal in multimodal machine learning ; images and texts are mapped to the same latent space which allows for comparison between objects of different modalities (Weston et al. 2010; J. Wu et al. 2017; Vukotic et al. 2018). Having a multimodal shared space has several applications as with the *cross-modal retrieval* task, which aims at retrieving relevant items in modality B with respect to a query in modality A. The *Image-retrieval* application is a common example where one seeks to retrieve relevant images given a textual query. Moreover, classification approaches with a large number of classes (above 1000) commonly use a multimodal shared space to represent the labels in the same space as the processed inputs; this allows image annotation systems to scale both in time and memory (Weston et al. 2010). Furthermore, as detailed in (Weston et al. 2010), using a multimodal shared space leads to more interpretable image annotation systems as, for example, synonyms or similar annotations are closely embedded.

There are several approaches to build a multimodal shared space, and we only describe here techniques trying to map one space to the other. Approaches that specifically aim at enriching textual space with visual information are detailed in Section 2.2. In the paragraphs below, we distinguish methods that compute a global alignment from approaches that use local metric learning rules.

Global alignment methods Given two mono-modal manifolds, *e.g.* one textual and one visual, global alignment methods aim to learn two mappings from each mono-modal manifold to a joint space, such that semantically similar regions across modalities are embedded closely in the shared space.

The first works that align image and text are based on the Canonical Correlation Analysis (CCA) (Hardoon et al. 2004). CCA aims at finding linear projections that maximize the correlation between pairs of items of textual and visual modalities (Silberer et al. 2012; Gong et al. 2014). Formally, given two data matrices $T \in \mathbb{R}^{N,d_t}$ and $V \in \mathbb{R}^{N,d_v}$, CCA seeks two vectors $a \in \mathbb{R}^{d_t}$ and $b \in \mathbb{R}^{d_v}$ such that the canonical correlation of $Ta^\top$ and $Vb^\top$ is maximized. Random variables $Ta^\top$ and $Vb^\top$ are called the first pair of canonical variables. The process can be iterated to find the second pair of canonical variables, $a' \in \mathbb{R}^{d_t}$ and $b' \in \mathbb{R}^{d_v}$, which maximize the canonical correlation of $Ta'^\top$ with $Vb'^\top$, subject to the constraint that they are uncorrelated with the first pair of canonical variables. After $n \leq \min(d_v, d_t)$ iterations, canonical projection matrices are thus obtained: $A \in \mathbb{R}^{n\times d_t}$ and $B \in \mathbb{R}^{n\times d_v}$ and the original data can be projected to a latent space of dimension n with $TA^\top$ and $VB^\top$. CCA allows to reduce the dimensionality of the linguistic and visual representations — which is desirable given the size visual representations can have — such that the important interactions between them are preserved.

CCA have be used to learn multimodal representations, which are formed by concatenating the transformed textual and visual embeddings Silberer et al. 2012;

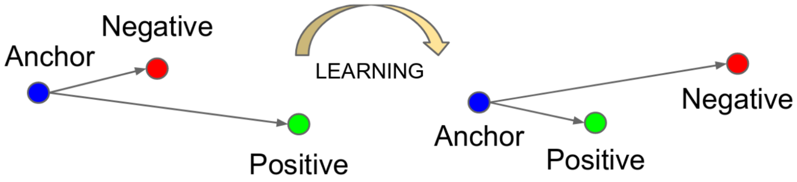


Figure 2.6 – **Illustration of the *triplet loss*.** Image taken from (Schroff et al. 2015).

Hill et al. 2014b; Gong et al. 2014; Plummer et al. 2017. Extensions of CCA include the use of kernelized version of CCA (Bach et al. 2002; Hill et al. 2014b), and the use of non-linear projections in the deep CCA (Andrew et al. 2013; Klein et al. 2015).

Local metric learning methods Unlike global alignment methods, local metric learning methods use local rules and updates to build a multimodal shared space. Under this setting, the problem of learning the shared space is cast as a ranking problem where matching elements should be embedded closer that non-matching elements.

A *pair-wise loss*, also called *contrastive loss*, can be used, where the distance $d(x_1, x_2)$ between positive (*i.e.* matching, $y = 1$) pairs of inputs (x_1, x_2) is minimized and the distance between negative (*i.e.* non-matching, $y = 0$) pairs is maximized to avoid trivial collapsing of the shared space. This is usually implemented with a hinge-loss function (Hadsell et al. 2006):

$$\mathcal{L}_{pairwise} = \sum_{(y,x_1,x_2)} y d(x_1, x_2) + (1 - y) \lfloor \gamma - d(x_1, x_2) \rfloor_+^2 \tag{2.7}$$

The margin γ is used to tighten the constraint: when two inputs x_1 and x_2 do not match, then their distance $d(x_1, x_2)$ should be bigger than the margin γ. This approach is also called *metric learning* as the objective learns the distance d between elements (Xing et al. 2002).

Besides pair-wise losses, *triplet losses* are used when no negative evidence is present. Triplet losses have the particularity to share an element between positive and negative pairs, this element is called the *anchor* (Zhu et al. 2015). The triplet loss enforces the distance $d(a, p)$ between an anchor a and a positive element p to be smaller, within a margin γ, than the distance $d(a, n)$ between the same anchor a and a negative element n (see Figure 2.6 for an illustration):

$$\mathcal{L}_{triplet} = \sum_{(a,p,n)} \lfloor \gamma + d(a, p) - d(a, n) \rfloor_+ \tag{2.8}$$

For example, Socher et al. 2014 use a triplet loss to learn alignment between images and their captions. Triplet losses have shown to be more efficient than pairwise losses to learn a shared space (Weinberger et al. 2009; Carvalho et al. 2018).

2.1.4 Beyond traditional multimodal approaches

In this section, we have reviewed representation learning techniques that handle textual and visual modalities, either separately (Section 2.1.1 and Section 2.1.2) or jointly (Section 2.1.3), and we have discussed various tasks that can be tackled with such techniques. At this point, the main issue arises from the fact that both modalities are, by nature, very differently structured (*e.g.* words vs. pixels), which makes challenging the design of (1) multimodal fusion strategies and of (2) a shared cross-modal space.

However, a key element is missing in our consideration of the textual and visual modalities: not only are modalities different in the way information is encoded, but also they are different in the expressed semantic. This point is developed and detailed in Section 2.2.1. This motivates approaches that consider each of the textual and visual modality as complementary sources of information, whose combination allows more semantics to be learned and represented than when considering each modality separately.

In the remaining of this chapter, we present challenges and techniques that involve both the textual and visual modalities, beyond the simple association approach. In particular, we show how a modality can be leveraged to benefit the other modality. Specifically, we discuss how:

- visual data can be used to augment linguistic capacities, this is called *language grounding*. It is presented in Section 2.2.

- conversely, language can be used to augment the capacities of visual recognition models. This is presented in Section 2.3.

2.2 Grounding natural language in the visual world

Motivated by the fact that language lacks common-sense information and is a biased view of reality (see Section 2.2.1), we review in this section how visual information can be used along textual data to learn common-sense knowledge and to enhance linguistic representations (see Section 2.2.2).

2.2.1 Motivation: *human reporting bias*

Textual and visual modalities are trivially different by the way information is represented: language is composed of sequences of tokens (words) arranged

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	368,922	31,168
Inhaled	984,613	5,617	Exhaled	168,985	4,052
Breathed	725,034	41,215	Was on time	23,997	14

Table 2.2 – **Illustration of the Human Reporting Bias**. Count of the number of times that *A person may <x>*, in Teraword and Knext textual corpora. Reproduced from (Gordon et al. 2013).

in sentences, paragraphs, documents. . . , while images are composed of parallel channels (RGB), with pixel values spatially arranged in two dimensions.

Beyond these obvious differences, it is also commonly assumed that these modalities do not bear the same semantics:

- Visual data are direct depictions of the reality and are not subject to interpretation: views of objects and spatial organization of scenes in images are unequivocal.

- Conversely, language refers to high-level concepts, can be ambiguous, relies on context and background knowledge (*e.g.* common-sense), and consequently is a biased view of reality.

The gap between reality and its textual description is called the *human reporting bias* (Gordon et al. 2013). In particular, Gordon et al. 2013 state that the frequency with which people refer to things or actions in language does not correlate with real world frequencies. For example:

- The more expected something is, the less likely people are to convey it as the primary intent of an utterance.

- The more value people attach to something, the more likely they are to give information about it, even if the information is unsurprising,

- Conversely, even unusual facts are unlikely to be mentioned if they are trivial,

- There are fundamental kinds of lexical and world knowledge that are needed for understanding and infering what is not stated in text.

This discrepancy between the textual and visual modalities can be explained by the fact that, by nature, natural language is produced by humans and addressed to other humans, and language is thus disconnected from a concrete reality. For example, when someone talks or writes, he makes the underlying assumptions that the people to which the language is addressed know about the world: many implicit facts are not mentioned as they are obvious and taken for granted for both parties. This has also been theorized and investigated in (Grice 1975; Ahn et al. 2005). To illustrate this, Gordon et al. 2013 analyzed huge web corpora and found discrepancies between real-world and textual statistics, as shown in Table 2.2. For example, based on *n*-gram occurrences count, they observe that a person is 3 times more likely to get murdered that to inhale. The inductive approach to hold textual references as evidence thus fails.

Under the prism of NLP, Bruni et al. 2012 show that purely textual models learn representations that do not contain the typical color of common concrete objects ; they analyze a purely-textual DSM and find for example that the closest color (in terms of cosine similarity) in the semantic space of the word *sky* is *green*, or *blue* for a *violin*. These simple findings show that it is challenging to extract common-sense knowledge and to perform real-world reasoning from purely textual data.

In psychological research, studies reveal that the meaning of concepts is grounded in perception (Glenberg et al. 2002; Barsalou 2008). From a neuroscience perspective, there is ample evidence that linguistic and visual processing are coupled (Ferreira et al. 2007). For example, it has been shown that language can alter the processing of visual information at early stages (Boutonnet et al. 2015; Kok et al. 2014).

The existence of the *human reporting bias* gives a motivation to seek information in images to provide complementary knowledge to the one extracted from text, *i.e.* to *ground* language in the visual world (Baroni 2016).

Common-sense. As an initial example about how leveraging images can complement language, Bagherinezhad et al. 2016 seek to reason about sizes of common objects, *e.g.* answering questions such as *what is the size of an elephant?*. Their model is based on the assumption that language gives access to absolute size of some objects, while images provide relative sizes of objects, *e.g.* when two objects are contained within the same image. They construct a *size-graph* where each vertex is a learnt (log-normal) distribution about an object and edges encode the relative size between two objects, as illustrated in Figure 2.7. They use textual information to provide absolute size of an object and visual information to provide relative size between two objects by comparing the depth-adjusted size of the two bounding boxed estimated with webly-supervised detectors (Divvala et al. 2014). Given observations from both modalities they maximize a likelihood to find optimal parameters of the distributions of the size for each object. Afterwards, they are able to perform graph inference to determine the size of an object.

This example underlines the complementary role that language and vision can play for common-sense approaches, and several works have proposed to learn *common-sense assertions*. Common-sense assertions model plausible relations between objects. They are typically modeled with triplets, *i.e.* relations of the form (*subject*, *predicate*, *object*). For example, to encode that a person rides a horse, the following triplet is used: (*man*, *ride*, *horse*). Traditionally, common-sense triplets

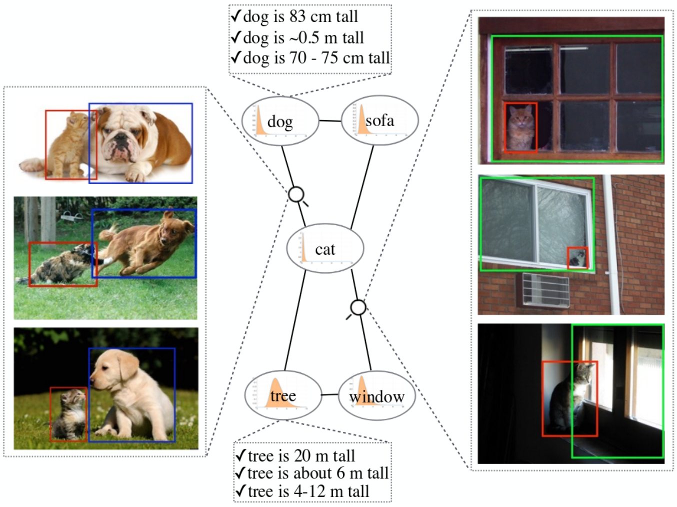
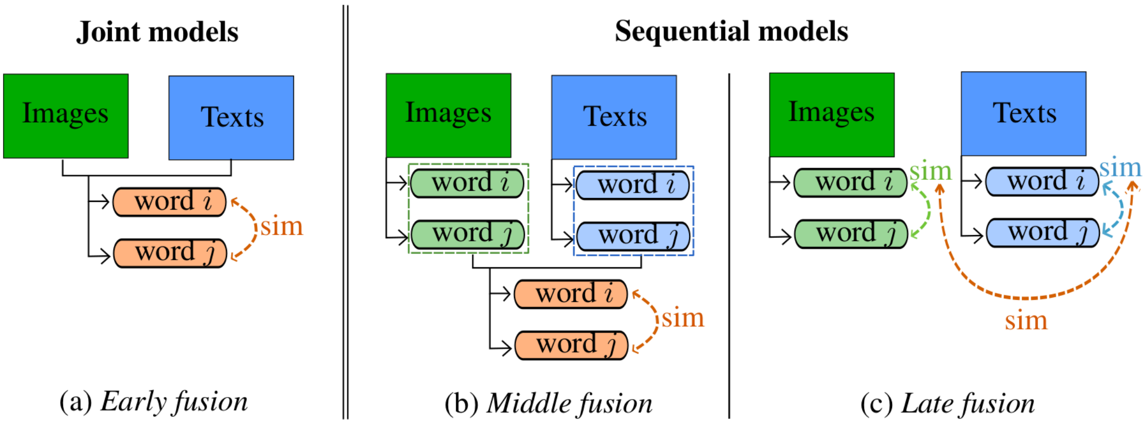


Figure 2.7 – **Exploiting textual and visual resources to reason about sizes of common objects**. In this example, language gives direct information about sizes of objects, while images gives information about the relative sizes of objects. Combining both direct and relative size knowledge allows to infer size of new objects. Illustration taken from (Bagherinezhad et al. 2016)

are collected and curated by humans and typically encoded within a KB (Lenat et al. 1990; Speer et al. 2017). Machine learning approaches were proposed to automatically extract common-sense triplets from textual resources (Downey et al. 2005; Vanderwende 2005; Etzioni et al. 2008; Carlson et al. 2010; Akbik et al. 2014; Jastrzebski et al. 2018) and more recently, several works have explored *multimodal* approaches to help capturing visual common-sense (Vedantam et al. 2015; Yatskar et al. 2016; F. Sadeghi et al. 2015).

For example, Vedantam et al. 2015 and X. Lin et al. 2015 propose to use abstract scenes (Zitnick et al. 2013) to learn common-sense; they argue that abstract scenes, despite not being photo-realistic, offer a media to generate semantically rich worlds where the annotations are known and do not suffer from imperfect predictions of object recognition and detection methods. The model of Vedantam et al. 2015 reasons about the plausibility of a common-sense triplet by measuring its similarity to both triplets extracted from texts, and to relations and nouns observed in abstract scenes. The final plausibility score is a linear combination of the textual and visual score. Beyond abstract scenes, Yatskar et al. 2016 seek to extract visual common-sense from natural images, which they argue contain information about spatial and functional properties of objects. They use a statistical approach to learn entailment rules between the spatial configuration of objects in an image (given by annotated bounding boxes) and the caption associated with the image. Their model relies on the computation of the Pointwise Mutual Information (PMI) of word pairs to estimate the evidence for each



(a) *Early fusion*

(b) *Middle fusion*

(c) *Late fusion*

Figure 2.8 – **Overview of multimodal fusion techniques.** Round-corner rectangles denote word embeddings. Green is related to images and blue to text, orange round-corner rectangles are multimodal embeddings built from textual and visual resources. "sim" stands for an example of an evaluation task, namely *word similarity*.

relationship triplet. Moreover, they also present a way to generalize the induced relations to related words using the semantic hierarchy of concepts in WordNet (Miller 1995).

Apart from these models which specifically target the common-sense assertion task, several approaches have been explored to leverage visual information for language understanding. They rely on incorporating visual information in a multimodal semantic space, *i.e.* learning multimodal general-purpose representations for linguistics units such as words or sentences.

2.2.2 Learning grounded linguistic representations

Representing language semantics is a long-standing problem for the natural language processing community as explained in Section 2.1.1, and to further improve traditional approaches towards that goal, we posit that leveraging visual information is crucial. We review methods that use images, along with textual data, to learn word (Section 2.2.2.1) and sentence (Section 2.2.2.2) representations.

2.2.2.1 Multimodal word representations

In a seminal paper, Nenov et al. 1988 propose to bind a word to its corresponding visual feature, such that word meaning gets grounded in the visual world. Since then, two main lines of multimodal DSM have been proposed to ground word semantics in the visual world: sequential models and joint models, as illustrated in Figure 2.8.

Sequential methods We suppose that entities can be identified both in images and text. Therefore, it is possible to learn textual representations $\{t_e\}$ with `GloVe` (Pennington et al. 2014) or `Word2Vec` (Mikolov et al. 2013), and visual representations $\{v_e\}$ from the aggregation (*e.g.* average or pooling) of activations of the penultimate layer of a pre-trained ConvNet on images of the entity e.

Sequential methods separately construct visual and textual word representations as explained above, and combine them using different techniques, *i.e.* through *middle fusion* or *late fusion*. Given separately learned representations in each modality, *i.e.* t_e the textual representation and v_e the visual representation for the entity e, middle fusion consists in merging them to form a multimodal vector m_e (see Figure 2.8 (b)). Several aggregation methods have been considered such as Concatenation (Kiela et al. 2014a), SVD (Bruni et al. 2012), CCA (Silberer et al. 2012), Weighted Gram Matrix Combination (Hill et al. 2014b) or the task-driven cross-modal mapping (Collell et al. 2017).

For example, the model of (Collell et al. 2017) is illustrated in Figure 2.9a. During training, a cross-modal function f is learned to map the textual representations t_e of the input entity e to its visual counterpart v_e, with a squared loss function. After training, the multimodal representation for the entity e is then obtained by concatenating ($\oplus$) the original textual representation t_e with the predicted visual vectors $f(t_e)$ (*imagined* vectors, as called by the authors):

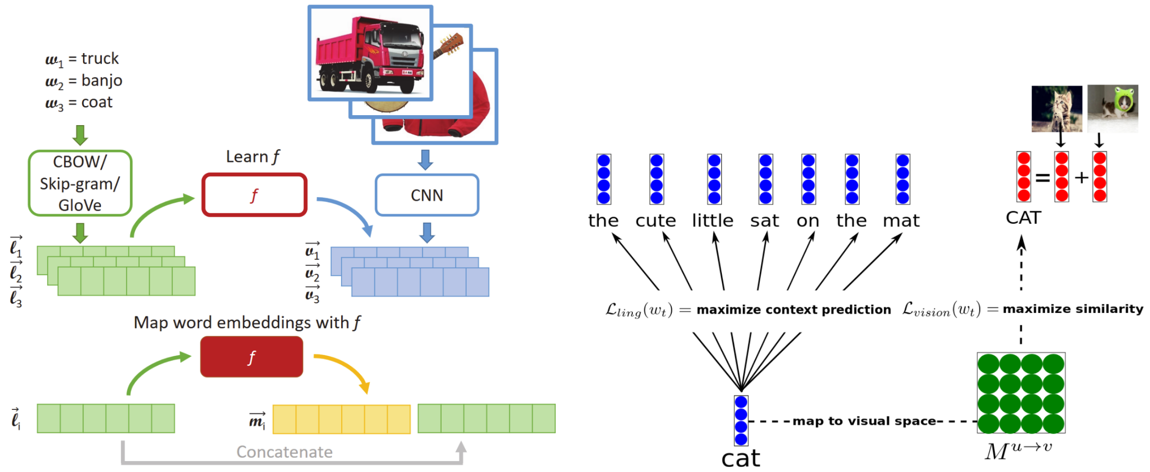
$$m_e = t_e \oplus f(t_e) \tag{2.9}$$

With this model, even abstract words, which do not have associated visual features, can benefit from grounding.

In late fusion (see Figure 2.8 (c)), word representations are computed separately for each modality. Their multimodal interactions only occur in a downstream task, as done in (Bruni et al. 2014), who use a simple linear combination of similarity scores respectively obtained from textual and visual data, to measure a global word similarity.

Joint methods Middle and late fusion models prevent potentially beneficial interactions during training between the different modalities. Joint models directly learn a joint representation from textual and visual inputs (Figure 2.8 (a)). This idea borrows from the way humans learn grounded meaning in semantics (Glenberg et al. 2002; Barsalou 2008). Some joint models require aligned texts and images. For example (Roller et al. 2013) use a Bayesian modeling approach based on the assumption that text and associated images are generated using a shared set of underlying latent topics and (Kottur et al. 2016) ground word representations into vision by trying to predict the abstract scene associated to a given sentence.

Extensions of the `skip-gram` algorithm (Section 2.1.1) have also been proposed, and they have the advantage of not relying on aligned text and images. For


w_1 = truck
w_2 = banjo
w_3 = coat
CBOW/ Skip-gram/ GloVe
Learn f
f
CNN
$\vec{\ell}_1$ $\vec{\ell}_2$ $\vec{\ell}_3$
$\vec{v}_1$ $\vec{v}_2$ $\vec{v}_3$
Map word embeddings with f
f
$\vec{\ell}_i$
$\vec{m}_i$
Concatenate
the cute little sat on the mat
CAT
$\mathcal{L}_{ling}(w_t)$ = maximize context prediction
$\mathcal{L}_{vision}(w_t)$ = maximize similarity
map to visual space
cat
$M^{u \to v}$

(a) **Example of sequential fusion.** Image taken from (Collell et al. 2017).

(b) **Example of joint fusion..** Image taken from (Lazaridou et al. 2015).

Figure 2.9 – **Example of multimodal fusion techniques for learning grounded word representations**

example, Hill et al. 2014a base their model on the assumption that the frequency of appearance of concrete concepts correlates with the likelihood of "experiencing" it in the world. Perceptual information for concrete concepts is then introduced to the model whenever that concept is encountered in the textual modality. Representations of concrete words are trained to predict surrounding words (as in the classical `skip-gram` model) and the perceptual features are feature-norms (McRae et al. 2005) that describe objects as a set of features (typical color, usage, etc.).

The work by Hill et al. 2014a was later followed by (Lazaridou et al. 2015) whose method is designed to use natural images instead of the handcrafted feature-norms. They force the representation of words for which they have images to be close to their visual (pre-trained) representation — their approach is illustrated in Figure 2.9b. More precisely, for any visual entity e, they assume that a visual vector v_e representing the entity is available. Typically, the visual vector $\{v_e\}$ is built from the average of activations obtained with a pre-trained `ResNet` applied on 100 images (taken from ImageNet) of the entity e. During training, along with a purely textual `skip-gram` loss, the similarity between the embedding t_e of the entity e and its visual appearance v_e is maximized in a max-margin framework:

$$\mathcal{L} = \sum_{e \in \mathcal{D}} \sum_{v^-} \max(0, \gamma - \cos(t_e, v_e) + \cos(t_e, v^-)) \tag{2.10}$$

where γ is the margin and v^- is the visual appearance of a "negative" object (randomly sampled over all objects, with uniform distribution). The visual representation v_e of each entity e is kept fixed throughout the optimization of the loss function.

word 1	*word 2*	similarity
cat	dog	0.76
stupid	dumb	0.91
advise	baseball	0.04
...	...	...

(a) **Similarity/Relatedness**

word	has_legs	can_fly	has_teeth	is_green	is_animal	...
monkey	1	0	1	0	1	...
crocodile	1	0	1	1	1	...
plane	0	1	0	0	0	...
...	...	...	...			

(b) **Feature-norms**

word	concreteness
dog	0.97
cloud	0.72
hope	0.15
...	...

(c) **Concreteness**

word 1	*word 2*	*word a*	*word b*
Paris	France	London	England
man	woman	king	queen
do	doing	eat	eating
...	...	...	...

(d) **Analogies**

Table 2.3 – **Word evaluation benchmarks**

Evaluation Once grounded word embeddings are learned, it is necessary to evaluate their quality. For this purpose, several *intrinsic* tasks exist to evaluate the quality of semantic representations of words (Bakarov 2018). These tasks use human-annotated benchmarks, and we report below the most popular evaluation tasks:[2]

- **Word similarity and relatedness benchmarks.** Semantic similarity (resp. relatedness) evaluates the similarity (resp. relatedness) degree of word pairs. We use several benchmarks which provide gold labels (*i.e.* human judgment scores) for word pairs: WordSim353 (Finkelstein et al. 2002), MEN (Bruni et al. 2014), SimLex-999 (Hill et al. 2015), SemSim and VisSim (Silberer et al. 2014). The *spearman* correlation is computed between the list of similarity scores given by the model (cosine-similarity between multimodal vectors) and the gold labels (as shown in Table 2.3a). The higher the correlation is, the more visual semantics are captured in the embeddings.

- **Feature norm prediction.** Collell et al. 2016 use the task of predicting feature-norms (*e.g.* 'is_red', 'can_fly') of objects given word representation to evaluate visual or textual-based representations (see Table 2.3b for an example). The evaluation dataset is an extract of the McRae dataset (McRae et al. 2005). There is a total of 43 characteristics grouped into 9 categories for 417 entities. A linear SVM classifier is trained and 5-fold validation scores are reported (F1

[2] We implemented these evaluation tasks (except the analogy prediction task) within a *plug-and-play* python library available here: `github.com/EloiZ/embedding_evaluation`

scores). In other words, this evaluation task amounts to assess if words that share similar characteristics (color, form, taste, purpose, ...) have the same direction in the representation space.

- **Abstractness / Concreteness prediction.** The USF norms (Nelson et al. 2004) give concreteness ratings for 3260 English words (see Table 2.3c). The goal is to predict this concreteness (regression) from word embeddings: with a word representation, we wish to know if it contains information that can be used to predict the concreteness rating of the associated word. In practice, we train an SVM with a Radial Basis Function (RBF) kernel to predict the gold concreteness rating from word embeddings. Reported scores are coefficients of determination (R^2).
- **Analogy prediction.** This task assesses whether word relationships correspond to geometrical relationships in the representation space (Mikolov et al. 2013). Given three words w_1, w_2 and w_a, the goal is to identify the word w_b such that the relation between w_a and w_b is the same as the one between w_1 and w_2. These relations can be grammatical, ontological, . . . (see Table 2.3d). There is no obvious choices to find the correct word w_b, but authors have considered methods based on arithmetic operations in vector space (addition and multiplication of cosine similarities) (Mikolov et al. 2013), or methods which additionally take the direction of the resulting vector into account in the evaluation (Levy et al. 2014b).

The benchmarks presented above are widely used for word embedding evaluation, however, they are only a proxy to evaluate the semantics contained in the vector space. This comes with the limitation that intrinsic evaluation scores do not necessarily correlate with downstream performances on real-world tasks (Faruqui et al. 2016; Bakarov 2018). As most real-world tasks are sentence-based, we describe such benchmarks in Section 2.2.2.2.

2.2.2.2 Multimodal sentence representation

While the literature is abundant about learning grounded *word* representations, there exist far less methods to learn grounded *sentence* representations. We are aware of two works which learn sentence representations, and both of them leverage images aligned with sentences in captioning datasets.

Chrupala et al. 2015 propose the `IMAGINET` model where two sentence encoders share word embeddings: a first GRU encoder learns a language model objective while the other one is trained to predict the visual features associated with a sentence. In this case, visual features correspond to the activations given by a pre-trained ConvNet applied on the image associated to the sentence. After training, the first encoder is kept and visual information of the second encoder has been transferred to the shared word representations.

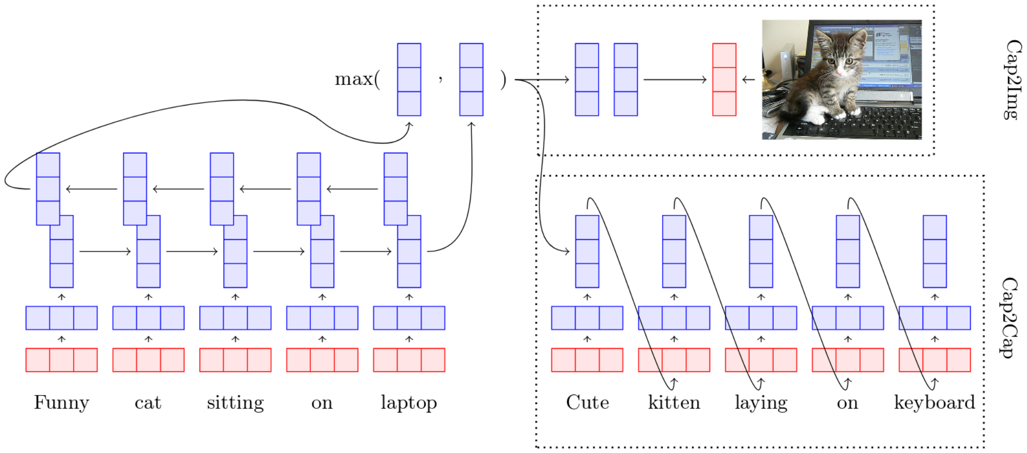


Figure 2.10 – **Learning visually grounded sentence representations**. Illustration taken from (Kiela et al. 2018)

The model of Kiela et al. 2018 is close to IMAGINET and additionally hypothesizes that associated captions ground the meaning of a sentence. They consider a bidirectional LSTM sentence encoder f_θ, parameterized by θ, which is trained according to two complementary objectives (an illustration is given in Figure 2.10):

On the one hand, the Cap2Img objective incorporates visual semantics into sentence representations by training sentence representations to reconstruct visual representations of the respective corresponding images using a triplet ranking loss. Given a captioning dataset $\mathcal{D} = (I, S)$ consisting of images, where each image I is associated with a caption S, the objective of the Cap2Img model is:

$$\mathcal{L}_{\text{Cap2Img}}(\theta) = \sum_{(I,S)\in\mathcal{D}} g_{\text{rank}}(i, f_\theta(S)) + g_{\text{rank}}(f_\theta(S), i) \tag{2.11}$$

where i is the latent representation of the image I, in this case, activation features obtained at the penultimate layer of a pre-trained ResNet , and where g_{rank} is a triplet loss, defined in Section 2.1.3.2:

$$g_{\text{rank}}(a, b) = \sum_n \lfloor \gamma - \cos(a, b) + \cos(a, n) \rfloor_+ \tag{2.12}$$

where γ is the margin, and n is randomly sampled from the set of negative elements.

On the other hand, the Cap2Cap objective enforces sentence representations to contain information about related sentences that describe the same image. Given two sentences S and S', which describe the same image, words composing S' are $(w_1, w_2, \ldots, w_n) = S'$. The encoder-decoder model is used: the sentence S'

needs to be decoded given the encoded representation $f_\theta(S)$ of the sentence S. The training objective of the Cap2Cap model is thus:

$$\mathcal{L}_{\text{Cap2Cap}}(\theta) = - \sum_{(S,S') \in \mathcal{D}} \sum_{i=1}^{n} \log p(w_i \mid f_\theta(S), w_1, \ldots, w_{i-1}) \tag{2.13}$$

where the probability p is parameterized with a softmax function:

$$p(w_i = k \mid f_\theta(S), w_1, \ldots, w_{i-1}) = \frac{e^{\langle k, h_i \rangle}}{\sum_{w} e^{\langle w, h_i \rangle}} \tag{2.14}$$

where, $\langle . \rangle$ denotes the cosine similarity, h_i is the representation of the decoder at step i, and the sum in the denominator ranges over all words of the dictionary.

The final sentence representations are obtained by concatenating (1) purely-textual SkipThought representations, and (2) grounded sentence vectors obtained with the Cap2Cap or Cap2Img (or both).

Evaluation of sentence representations Like in the case of words, several tasks and benchmarks exist to evaluate sentence representations. We list below two widely used tasks, along with their corresponding benchmarks:

- **Semantic relatedness.** Semantic similarity benchmarks such as Semantic Textual Similarity (STS) (D. M. Cer et al. 2017) and Sentences Involving Compositional Knowledge (SICK) (Marelli et al. 2014a), consist of pairs of sentences that are associated with human-labeled similarity scores. STS is subdivided into three textual sources: *Captions* contains concrete sentences describing daily-life actions, whereas the others contain more abstract sentences: news headlines in *News* and posts from users forum in *Forum*. Spearman correlations are measured between the cosine similarity of learned sentence embeddings and human-labeled scores.

- **Classification benchmarks.** Several downstream classification tasks are commonly used, and they are implemented in the SentEval library[3] (Conneau et al. 2018). The tasks are the following: opinion polarity (Multi-Perspective Question Answering (MPQA)) (Wiebe et al. 2005), Movie Review (MR) (Pang et al. 2005), subjectivity/objectivity classification (SUBJ) (Pang et al. 2004), Customer Reviews (CR) (Hu et al. 2004), binary sentiment analysis on Stanford Sentiment Treebank (SST) (Socher et al. 2013), paraphrase identification (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.

[3] github.com/facebookresearch/SentEval

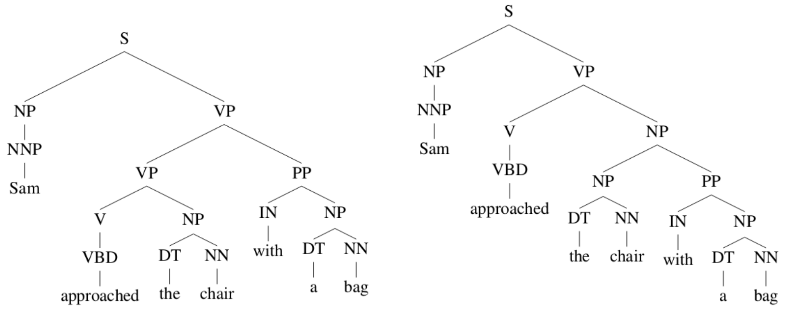
Fill-in-the-blank	Mike is having lunch when he sees a bear. ________ A. Mike orders a pizza B. Mike hugs the bear C. Bears are mammals D. Mike tries to hide
Visual paraphrasing	Are these two descriptions describing the same scene? 1. Mike had his baseball bat at the park. Jenny was going to throw her pie at Mike. Mike was upset he didn't want Jenny to hit him with a pie. 2. Mike is holding a bat. Jenny is very angry. Jenny is holding a pie.
Language desambiguation	"Sam approached the chair with a bag"  S NP VP NNP Sam VP PP V NP IN NP VBD DT NN with DT NN approached the chair a bag S NP VP NNP Sam V NP VBD approached NP PP DT NN IN NP the chair with DT NN a bag

Table 2.4 – **Linguistic common-sense tasks.** Reproduced from (X. Lin et al. 2015; Berzak et al. 2015)

NLP evaluation tasks directly targeting visual information Beyond sentences, we present below a list of purely-textual tasks which have been shown to benefit from auxiliary visual data; some examples are illustrated in Table 2.4:

- *Lexical Preference*: This task consists in determining the plausible noun arguments for particular verb predicates. Bergsma et al. 2011b show that using visual information helps making better linguistic decision for this task. For each noun, visual features are extracted from corresponding web images, and for each verb a visual classifier is learned to select salient visual features of its preferred arguments.

- *Prepositional Phrase Attachment Resolution*: This task is related to the *lexical preference* task. It consists in determining the correct attachment between prepositional phrases (Christie et al. 2017; Berzak et al. 2015).

- *Visual paraphrasing evaluation*: The aim of this task is to assess if two sentences are likely to describe the same underlying (unseen) scene (X. Lin et al. 2015; Kottur et al. 2016). This task can evaluate the capacity to detect visually similar phrases as explored in (Divvala et al. 2014) (*e.g.* a *grazing horse* is visually similar to a *eating horse*).

- *Fill-in-the-blank*: This task consists in predicting a plausible fit in a blank left in a story. This task is very challenging as it directly require common-sense. For example, as illustrated in Table 2.4, answering the question involves the implicit knowledge that bears might be dangerous animals, and that people usually stay away from dangerous animals. (X. Lin et al. 2015) use this task to evaluate their multimodal models.

- *Bilingual lexicon induction*: The task is related to the synonym detection task. The aim is to find words across language that share a common meaning. The underlying hypothesis is that words with similar meaning across language correspond to similar images. (Bergsma et al. 2011a) use SIFT representation for images and their model return the Nearest Neighbors over the cosine similarity. (Kiela et al. 2015b) improve their approach by using ConvNet features instead of the SIFT features. (Vulic et al. 2016) create a multi-lingual multimodal space using `Word2Vec` features in addition to the visual features returned by a ConvNet on images found by Google image search.

- *Multimodal machine translation*: The aim is to translate a sentence from a source language to a target language, when an image is given to illustrate the sentence. Caglayan et al. 2016; Caglayan et al. 2017 show that using images helps to produce improved translations, both for machine learning systems, and for humans. While this is a promising research direction, to the best of our knowledge, multimodal Neural Machine Translation (NMT) systems show similar performances with unimodal NMT systems on the text-only machine translation task. Transferring visual knowledge acquired on the multimodal task to the scenario where no images are provided remains an open challenge.

There is room left for improvement to incorporate visual information in NLP approaches. It has indeed not yet been proven that using visual information can help more "canonical" NLP tasks such as automatic summarization, open-domain question-answering and semantic role-labeling. We discuss this research direction in Section 6.2.

2.3 Visual understanding with natural language

We reviewed in Section 2.2 several approaches that leverage visual information to bring complementary information to language, in particular to improve NLP representations and perform common-sense reasoning. We now take the opposite approach, and raise the simple question:

How can language help computer vision?

Regarding this question, we detail in this section two orthogonal research directions:

1. language helps as a way to evaluate capacities of visual recognition models (Section 2.3.1),
2. language can help to augment visual understanding capacities, in particular when visual supervision is scarce (Section 2.3.2).

2.3.1 Evaluating visual models with natural language

Historically, the main goals in learning-oriented computer vision are learning to *classify*, *detect* or *segment* objects. Over the last decade, huge advances have been made regarding these tasks. For example, given enough training data, super-human performances can be obtained on classification task of the *ImageNet* yearly competition (Krizhevsky et al. 2012; Russakovsky et al. 2015).

Beyond these tasks, some fundamental questions remain:

- How to evaluate the global scene understanding of visual systems? In particular, how well does the system focus on most salient objects and understands the relationships between objects? We present below (Section 2.3.1.1) the captioning task which answers these questions.

- In addition to the previous questions, what are the reasoning capacities of visual systems? As an answer to this question, the VQA task emerged and we detail it in Section 2.3.1.2.

2.3.1.1 Evaluating global scene understanding: the case of *captioning*

Towards global scene understanding, the image *captioning* task has been proposed. Captioning is the process of generating textual description of an image (or a video). Based on the assumption that language can express high-level semantics, the purpose of this task is to evaluate the capacity of a visual recognition system to extract the semantics of a scene. In particular, caption generation models must be capable of:

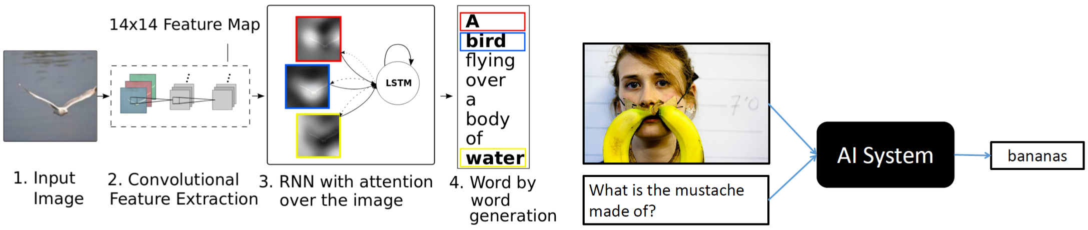
- recognizing objects and their attributes,

- recognizing relations between objects, spatial organization, actions and movements,

- selecting the salient pieces of information worth to be mentioned,

- expressing it in natural language.

Captioning models usually employ a *encoder-decoder* strategy (Kiros et al. 2014). Visual features are encoded in a latent space (usually with a pre-trained ConvNet)


14x14 Feature Map
LSTM
A
bird
flying
over
a
body
of
water
1. Input Image
2. Convolutional Feature Extraction
3. RNN with attention over the image
4. Word by word generation
AI System
bananas
What is the mustache made of?

(a) **Captioning task**. Model and illustration from (Xu et al. 2015)

(b) **VQA task**. Image taken from `visualqa.org`.

Figure 2.11 – **Visual tasks evaluated with natural language**

and then the visual representation is fed to a decoder network (usually a LSTM or GRU architecture) which sequentially generates a caption.

The use of attention, where the model learns to attend specific parts of image sequentially to generate a caption (illustrated in Figure 2.11a), has produced good results (Xu et al. 2015; Engilberge et al. 2018). Some works propose to optimize the metrics which are used to test the models, such as BiLingual Evaluation Understudy (BLEU) and Metric for Evaluation of Translation with Explicit ORdering (METEOR), using reinforcement learning algorithms as these metrics are not differentiable (Ranzato et al. 2016; Z. Ren et al. 2017).

A variant of the image captioning task has been proposed as the *Dense captioning* task, where the aim is to caption image regions and not only the whole scene (Johnson et al. 2016). Recently, Feng et al. 2018 explore the *unsupervised image captioning* task, where no alignment supervision is available between captions and images.

The captioning task comes with some limitations and the task is not fully sufficient to assess visual recognition and reasoning capacities. On the one hand, it has been shown that captioning systems exhibit biases from the training set: generated sentences usually corresponds to scene configurations that are seen in the training set, and poorly generalize to new ones (Hendricks et al. 2018). On the other hand, evaluating natural language outputs is known to be a thorny problem (Novikova et al. 2017), and it is widely known that higher scores on BLEU and METEOR metrics does not always imply an improved quality for captions in terms of human judgement, as many different sentences could be potential captions for a given scene.

2.3.1.2 Evaluating visual reasoning: the case of VQA

Regarding the limitations of the captioning task to evaluate visual recognition systems, the VQA task has recently been proposed (Malinowski et al. 2014). Given an image, the VQA task consists in answering a natural language question about the image (see illustration in Figure 2.11b). VQA is one of the most challenging task at the intersection of NLP and Computer Vision (CV) as high-level understanding of the image is needed and reasoning capacities with natural language is required to attend specific image parts and answer questions. In particular, VQA is a playground task for the following challenges:

1. designing efficient **multimodal fusion strategies** (discussed in Section 2.1.3.1). One core component of the VQA systems is the fusion model that merges linguistic and visual information from (1) the question, (2) the image, and (3) the answer (language) (Z. Yu et al. 2017; Ben-younes et al. 2017).
2. studying **reasoning capacities** of models. Some works specifically study visual reasoning capacities, such as counting, and inferring spatial relations and logical operations. For instance, the Compositional Language and Elementary Visual Reasoning diagnostics (CLEVR) dataset (Johnson et al. 2017) has been proposed to study the capacity of systems to understand complex queries (*e.g. how many cylinders are either green or smaller than the red ball?*), while keeping a simple visual environment (few objects, colors, and sizes).
3. studying **biases in the data**. A major challenge in VQA is to prevent models to overfit data biases (which is hard to avoid and quantify in the captioning task). For example, purely textual models usually reach good prediction scores, while they ignore the input image (Ramakrishnan et al. 2018). Recently, the VQA-2 dataset has been conceived where a question is asked on similar images but different answers are expected (Goyal et al. 2017). A good visual module is thus needed to learning reliable systems on this dataset.
4. designing **interpretable models**, *i.e.* models which can answer questions such as: Why did the model produced this answer? This long-term goal, which applies to the vast majority of machine learning tasks, has a prominent importance in the VQA setting. Towards this goal, attention modules can be used to visualize decisions made by the model (Cadène et al. 2019).

Towards the goal of learning and evaluating capacities of visual reasoning, some other tasks have been proposed, sharing similar ideas with the VQA task. For example, in the *Visual Dialog* task (Das et al. 2017; Mostafazadeh et al. 2017), an agent is required to hold a meaningful conversation with a human about a given image. Besides, in the *Visual Object Discovery through Visual Dialogue* task (Vries et al. 2017), a user seeks to locate an unknown object in a scene by asking a sequence of natural language questions to an agent.

2.3.2 Augmenting visual understanding systems with natural language

In the previous section, we saw examples where language can be used at the benefit of visual systems, as a way to evaluate global understanding and visual

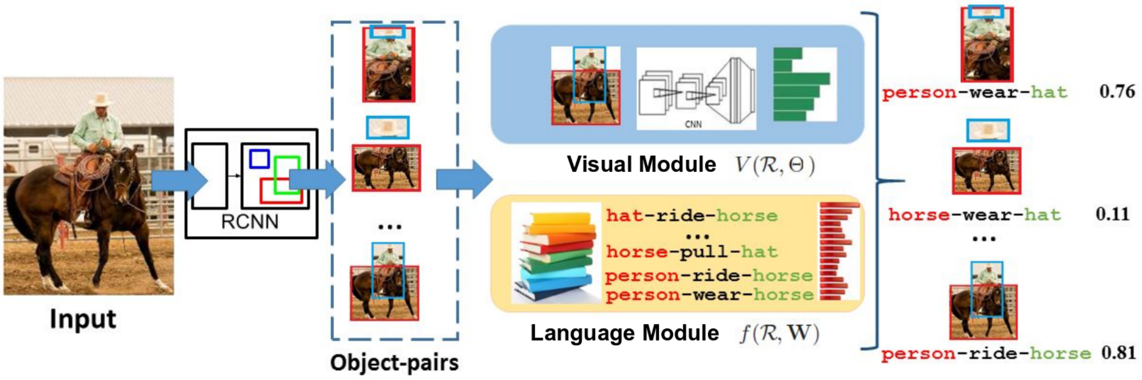


Figure 2.12 – **VRD task, with language priors**. Illustration from (C. Lu et al. 2016)

reasoning capacities. We now argue that language, *e.g.* in the form of semantic representations, can be used to improve capacities and performances of visual recognition systems.

Leveraging auxiliary linguistic knowledge for computer vision tasks is a promising research direction, as visual data is expensive to annotate while linguistic knowledge is readily available within huge amount of text corpora such as *Wikipedia*. Based on the assumption that language does contain visual information about the objects, language priors and word semantics can be used when visual supervision is a limiting factor, which is the case in the Visual Relationship Detection (VRD), and Zero-Shot Learning (ZSL) tasks (detailed below). Besides, recent studies show that using linguistic knowledge yields bigger improvements to visual models than increasing the size of the training dataset (R. Yu et al. 2017).

2.3.2.1 Using language priors for Visual Relationship Detection (VRD)

Like in the case of common-sense triplets, a *visual relation* is a triplet *(subject, predicate, object)*. The VRD task consists in detecting visual relations, *i.e.* finding pairs of objects and classifying each pair into a predicate that explains the relationship between the two objects (M. A. Sadeghi et al. 2011; C. Lu et al. 2016). Jung et al. 2019 enumerate three major difficulties encountered with the VRD task:

1. intra-class variance, where an object or the predicate is visually different depending on the other elements of the triplet. For example, “fishing” in the triplets (bear, fishing, salmon) and (man, eating, salmon) are visually very different (F. Sadeghi et al. 2015).

2. long-tail distribution of triplets, as many triplets are rarely or never seen. This makes it very hard to generalize to unseen triplets during training and it motivates to separately recognize objects and predicates rather direct predict the triplet as a visual phrase.

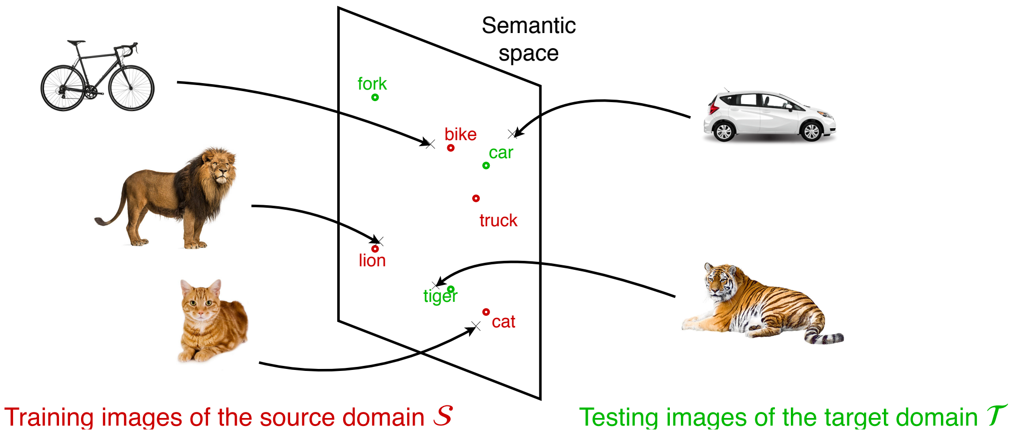

Semantic space
fork
bike
car
truck
lion
tiger
cat
Training images of the source domain $\mathcal{S}$
Testing images of the target domain $\mathcal{T}$

Figure 2.13 – **Classical ZSL model, without context**. Red color corresponds to the source domain and green to the target domain. ZSL models are trained using images of the source domain (on the left) and a mapping from the visual space to the semantic space is learned. At inference, testing images of classes of the target domain are projected in the semantic space.

3. class overlapping, where predicates can have very similar meanings (under vs. below), (close to vs. next to) but the model will be penalized for not giving the exact predicate.

Several approaches have been proposed to address these issues, and some of them propose to use language priors for the VRD task. The global idea is to *distill* external linguistic knowledge in the visual model. Distillation is an idea proposed by Geoffrey E. Hinton et al. 2015, where knowledge from a model A is transferred to another model B by teaching B to predict A's outputs; this can be viewed as a way to regularize the visual model with linguistic statistics.

For example, (R. Yu et al. 2017) mine linguistic knowledge from both the VRD training dataset and external data sources such as *Wikipedia*, and this knowledge is distilled in a teacher-student knowledge distillation framework. In (C. Lu et al. 2016), relations are projected to a linguistic space where representations of similar relationships are optimized to be close to one another. This allows the model to give non null probabilities to unseen triplets, if they are likely in language, and to account for the class overlapping problem mentioned above. An illustration of their approach is depicted in Figure 2.12.

2.3.2.2 Zero-Shot Learning (ZSL) for object recognition

While state-of-the-art image classification models (Zoph et al. 2017; Real et al. 2018) restrict their predictions to a finite set of predefined classes, ZSL bypasses this important limitation by transferring knowledge acquired from seen classes (*source domain*) to unseen classes (*target domain)*. Generalization is made possible through the medium of a common semantic space where all classes from both source and target domains are represented by vectors called *semantic representations*.

Historically, the first semantic representations that were used were handcrafted attributes (Farhadi et al. 2009; Parikh et al. 2011; Mensink et al. 2012; Lampert et al. 2014). In these works, the attributes of a given image are determined and the class with the most similar attributes is predicted. Most methods represent class labels with binary vectors of visual features (e.g, 'IsBlack','HasClaws') (Lampert et al. 2009; Liu et al. 2011; Y. Fu et al. 2014; Lampert et al. 2014). However, attribute-based methods do not scale efficiently since the attribute ontology is often domain-specific and has to be built manually.

To cope with this limitation, more recent ZSL works rely on distributed semantic representations learned from textual datasets such as Wikipedia, using DSM (Mikolov et al. 2013; Pennington et al. 2014; Peters et al. 2018). These models are based on the *distributional hypothesis* (Harris 1954), which states that textual items with similar contexts in text corpora tend to have similar meanings. This is of particular interest in ZSL: all object classes (from both source and target domains) are embedded into the same continuous vector space based on their textual context, which is a rich source of semantic information. Some models directly aggregate textual representations of class labels and the predictions of a ConvNet (Norouzi et al. 2013), whereas others learn a cross-modal mapping between image representations (given by a ConvNet) and pre-learned semantic embeddings (Akata et al. 2015; Bucher et al. 2016). At inference, the predicted class of a given image is the nearest neighbor in the semantic embedding space. The cross-modal mapping is linear in most of ZSL works (Palatucci et al. 2009; Romera-Paredes et al. 2015; Akata et al. 2016; Qiao et al. 2016). Among these works, the DeViSE model (Frome et al. 2013) uses a max-margin ranking objective to learn a cross-modal projection f between the image $\mathcal{V}$ and the semantic representation w_i of its label i:

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

where j is negative, uniformly-sampled label, w_j is its representation, and γ is the margin (hyper-parameter).

Several models have built upon DeViSE , by learning non-linear mappings between the visual and textual modalities (Ba et al. 2015; Xian et al. 2016), or by using a common multimodal space to embed both images and object classes (Z. Fu et al. 2015; Long et al. 2017). An illustration of the classical approach, that ignores visual context, for the ZSL task is given in Figure 2.13.

In Chapter 5, we extend the DeViSE model in two directions: by additionally leveraging visual context, and by reformulating it as a probabilistic model that allows coping with an imbalanced class distribution.

Very recently, an extension of the zero-shot object recognition has been proposed, where the aim is both to detect and classify an object that has never been seen, *i.e.* the *zero-shot object detection* task (Bansal et al. 2018; Demirel et al. 2018).

2.4 Positioning

In Chapter 3, we present new hypotheses, and a corresponding model, to learn *grounded word representations*. We build on previous works, such as the `skip-gram` algorithm (Section 2.1.1.3) and approaches leveraging visual information (Section 2.2.2.1), but we focus on taking into account the visual context of objects, and their spatial organization in images.

In Chapter 4, we present a model to learn *grounded sentence representations*. Building on previous works (Section 2.2.2.1 and Section 2.2.2.2), we argue that the use of cross-modal projections over-constraints the learned space in the case of sentences. Instead, we introduce two complementary objectives and propose to incorporate visual semantics within an intermediate space.

In Chapter 5, we present a model for *zero-shot object recognition*. Building on previous works, and on the `DeViSE` model in particular (Section 2.3.2.2), which assumes that information about the objects' appearance is contained in semantic representation, we formulate an additional hypothesis which is two-fold: information about possible visual environment and visual occurrence likelihood is contained in semantic representation. To evaluate these hypotheses, we design the *context-aware zero-shot recognition* task and propose and adequate model.

Chapter 3

GROUNDING LANGUAGE IN THE VISUAL WORLD: THE CASE OF WORDS

Contents

Chapter abstract

Representing the semantics of words is a long-standing problem for the Natural Language Processing (NLP) community. Most methods compute word semantics given their textual context in large corpora, and, more recently, researchers attempted to integrate perceptual and visual features. Most of these works only consider the visual appearance of objects to enhance word representations but they ignore the visual environment and context in which objects appear. In this chapter, we propose to unify text-based techniques with vision-based techniques by simultaneously leveraging textual and visual context to learn multimodal word embeddings. We explore various choices to capture the visual

context and present an end-to-end model to integrate visual context elements. We provide experiments and extensive analysis of the obtained results.

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

- Eloi Zablocki, Benjamin Piwowarski, Laure Soulier, and Patrick Gallinari (2018). "Learning Multi-Modal Word Representation Grounded in Visual Context". In: *AAAI 2018*.

3.1 Introduction

To further improve the quality of word representation, leveraging multimodal information is crucial. Indeed, psychological studies have given pieces of evidence that the meaning of words is grounded in perception (Glenberg et al. 2002; Barsalou 2008) and Gordon et al. 2013 report a bias between what is said in texts and what can be seen in images (detailed in Section 2.2.1). These observations outline the complementary roles of images and texts and bring new perspectives to multimodal approaches bridging textual information with visual ones to improve natural language processing tasks (Hill et al. 2014a; Lazaridou et al. 2015). Besides, it is worth mentioning that this has become possible thanks to the exploitation of significant advances in computer vision (Section 2.1.2).

Multimodal representation learning models have been proposed to enhance word representations using either sequential (Kiela et al. 2014b; Bruni et al. 2014) or joint fusion techniques (Hill et al. 2014a; Lazaridou et al. 2015), as detailed in Section 2.2.2.1. However, most of these works ignore the *visual context* of objects. We posit that learning representations of contexts in different modalities should be a key component of multimodal Distributional Semantic Model (DSM). The importance of context is illustrated in a simple example (Figure 3.1). From an image of an apple on a black background, we can see its color, its texture and shape. From its context, *e.g.* growing on a tree, we can infer the relative size of apples with respect to the tree leaves, and that apples are fruits that grow on trees. If there is someone that is eating the apple, we can infer that apples are edible, and so on. From this example, we understand why exploiting the visual surroundings and context of objects might be useful to grasp the semantics of words.

In this work, we propose a multimodal model for learning word representation, leveraging contexts in different modalities, namely texts and images. Our contribution is threefold:

- We propose and experiment with various definitions of what visual context is (Section 3.3.1) – this has never been taken into account to the best of our knowledge in such models;

- We propose a multimodal context-driven model to jointly learn representations from textual and visual modalities, where both modalities influence media-independent word embeddings (Section 3.3.2). One further strength of the model is that it does not require aligned images and text (*i.e.* images with captions);

- We present a thorough analysis of the obtained results to determine the influence of the visual modality on the learned multimodal embeddings (Section 3.4 and Section 3.5) by experimenting with a set of word classification tasks.

3.2 Visual contexts and research questions

3.2.1 Using and modeling visual contexts

Several of the works presented in Section 2.2.2.1 (Learning multimodal word representations) use the visual modality to constrain the textual representation to be close to the visual representation of the object. Such a strategy has two drawbacks. First, there is an asymmetry in the consideration of the modalities: text defines a semantic context for each word — its surrounding words — while images are used to gather visual information about the object. Second, it does not use the fact that the context in which objects appear is informative and complementary to textual inputs to improve word representation. Indeed, this fact is suggested by several works, such as (Bruni et al. 2012) who propose a middle fusion approach where a visual embedding is built by factorizing the matrix counting visual words in images. This is the first attempt to apply the distributional hypothesis to images: *Semantically similar objects will tend to occur in similar environments in images*. Through their experiments, they come to the conclusion that the appearance of the surrounding objects is more informative for semantics than the appearance of the object itself. In comparison to the model we present in this chapter, their work does not propose to jointly learn embeddings from both visual and textual context.

This statement is strengthened with observations in (Roller et al. 2013) and (Bruni et al. 2014). The former proposes a Latent Dirichlet Allocation (LDA) model. The latter uses a count-based technique to learn multimodal word embedding by leveraging both visual and textual contexts. First, they build target-context count matrices for text (count of co-occurrence patterns with contexts) and images, using bag-of-visual words to represent images. They concatenate both matrices and perform rank reduction with Singular Value Decomposition (SVD). They then split back the matrix into the original matrices which are called *smoothed* as the global matrix has been reduced. A smoothed text and a smoothed image matrix are thus obtained and fusion is considered at the feature level (middle fusion) or scoring level (late fusion). However, they use a "count-base" method which does not learn representation for contexts and performs poorly on semantic tasks. Moreover, their approach uses bags of visual words representation for images.

In addition to the identification of entities and their context, rich spatial information is present if objects can be located in the image. Bruni et al. 2014 propose to use spatial information for contexts by dividing the image in 4x4= 16 bins and performing the visual word extraction and counting pipeline separately for each region. However, when it comes to learning representations for words, exploiting spatiality is challenging and still largely under-explored.

3.2.2 Research questions

From reviewing the literature about learning multimodal word representation (Section 2.2.2.1), we observe three main issues with current multimodal DSM for which there are no consensual answers:

- Text and images are very different by nature (Gordon et al. 2013). A sentence has a linear structure with a sequence of tokens (words) while an image has spatially-organized quantifiable information (pixel values). In the `skip-gram` model, choosing surrounding words to be the context is a natural choice for a text, however, in images, it is not clear what should be used as context to learn semantically rich representations for objects (Roller et al. 2013; Bruni et al. 2014)).

- Several multimodal fusion methods exist, but none of the models presented above is significantly better than the others, and the question to know how to build a multimodal framework has no obvious answer, especially when the alignment between texts and images is missing.

- Evaluation tasks to assess the quality of word embeddings are inherently biased (Faruqui et al. 2016), and it is hard to examine in depth the contribution brought by the visual modality (Collell et al. 2016).

In contrast to other works in learning multimodal word representations, we posit that exploiting the visual context enhances the learned representation of words. This assumption draws us to consider images of complex scenes containing many objects. Indeed, images of a single object give very little information about the object, how it is used for, where it can be found and so on. On the contrary, an image showing an object in its environment, being used or interacting with other objects, is much more informative thanks to the surrounding context. Accordingly, we address the following research questions, also illustrated in Figure 3.1:

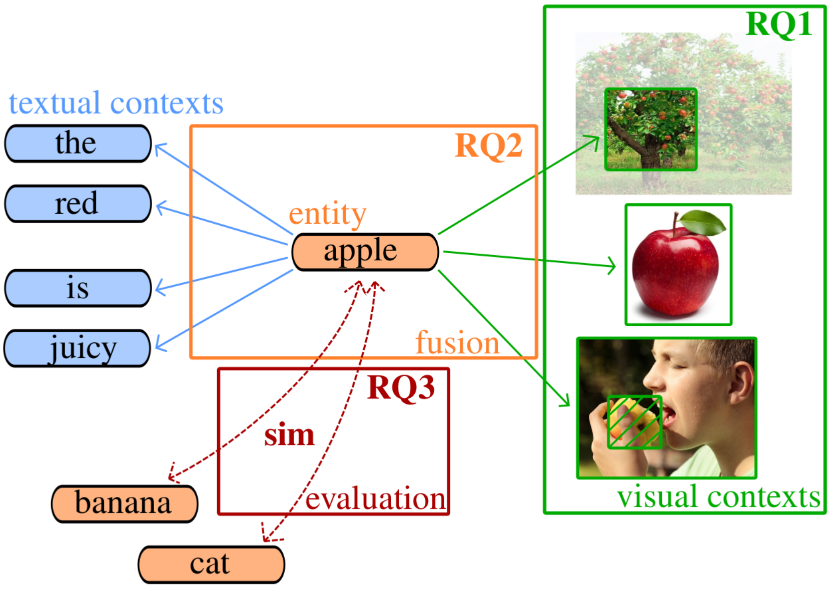


Figure 3.1 – **Illustration of the approach and underlying research questions.** Research Question (RQ)1 concerns using visual contexts for the visual part of the model, RQ2 is about the integration of the visual part with the text model and RQ3 deals with the evaluation of the embeddings.

- **RQ1**: In images, what can be used to learn semantic representations for objects? In particular, does context can capture some of the semantic of a word/entity? Note that in this work, we consider that the set of entities is the subset of the set of words that correspond to objects in images.
- **RQ2**: How can we naturally integrate a visual model with a text-based model to form a multimodal DSM?
- **RQ3**: How can we evaluate and examine the contribution given by the visual modality in the final word embeddings?

3.3 Model: Learning MultiModal Context-Driven Word Representations

We present here a multimodal DSM model leveraging both visual and textual contexts of words, relying on a multimodal distributional hypothesis. To do so, we formalize a definition of visual context (Section 3.3.1.1) and propose experiments to select appropriate visual context elements (RQ1, Section 3.5).

We then introduce our multimodal joint model based on the `skip-gram` framework (Mikolov et al. 2013) (RQ2, Section 3.3.2). The textual and visual parts of the model share the same word embeddings which are updated from both textual and visual inputs, with modality-specific contexts. One strength of our model relies on the fact that it does not require aligned data. Note that we assume that objects are already detected in images, which is not a strong assumption given the progress made in object detection.

3.3.1 Representation learning with visual contexts

In this section, we formalize what we name *visual contexts* and detail the choice of modeling that we propose. The different choices are illustrated in Figure 3.2.

3.3.1.1 Formalization.

Based on the original `skip-gram` algorithm that considers entities e (words) and their contexts $\mathcal{C}_e = \{c_1, ..., c_n\}$ (n surrounding words within a window centered on the entity), we translate in what follows the distributional hypothesis for images.

In our case, the contexts $\mathcal{C}_e$ are visual contexts. The choice for visual context elements $c \in \mathcal{C}_e$ does not need to correspond to a list of semantic entities, as shown by Levy et al. 2014a who propose a generalized `skip-gram` algorithm with arbitrary contexts. For instance, visual context elements can be the surrounding objects, low-level features such as the visual appearance, or also the localization of the surrounding objects with respect to the considered entity.

With this in mind, we define a function f_θ, parametrized by θ (learned), such that for any entity e and visual context element $c \in \mathcal{C}_e$, $f_\theta(c)$ is a vector of $\mathbb{R}^d$. These representations are then used in the negative-sampling loss:

$$\mathcal{L}_\mathrm{i} = -\sum_{e \in \mathcal{D}} \sum_{c \in \mathcal{C}_e} \left[\log \sigma(f_\theta(c)^\top t_e) + \sum_{c^-} \log \sigma(-f_\theta(c^-)^\top t_e) \right] \tag{3.1}$$

where $\mathcal{D}$ is the set of entities, t_e is the embedding associated with the entity/word e (learned), c^- is a negative context, and σ is the sigmoid function. This loss formulation is very close to the original `skip-gram` loss but integrates the learning of f_θ. The function f_θ projects the visual representation into the textual space and f_θ shares parameters (θ) for the computation of every context element.

3.3.1.2 Choice of modeling.

Given an entity e, we now propose different ways of modeling an instance of visual context elements $c \in \mathcal{C}_e$ and detail how to build and parametrize f_θ.

Surrounding objects (high-level context). An image I can be seen as a bag of objects: $I = \{o_1, o_2, ...\}$. This simple view gives high-level information about the environment in which objects occur. Given an entity $e = o_i$ (for some i) in an image, we define $\mathcal{C}_e = \{o_j, j \neq i\}$ as the set of all other objects that appear in the image. Then, each context $c = o_j \in \mathcal{C}_e$ is a surrounding object. We define $f_\theta(c) = V_c$ where $V \in \mathbb{R}^{M \times d}$ is a simple lookup table of embeddings for M objects, d the dimension of the representation space, and V_c the c^{th} row of this matrix.

Image patches (low-level context). At a coarser level, the set $\mathcal{C}_e$ of all visual context elements can be seen as image patches from the full image where entity e is masked out (with black pixels). We call this the *low-level context* since it directly uses pixel values from the surroundings of entities. Using low-level context is interesting because some objects can be left unidentified in images by current models. However, this requires a bigger and more complex model for processing the pixel values instead of the list of objects, and it is more difficult to extract meaningful information from pixel values. We suggest two possible choices to select $c \in \mathcal{C}_e$:

1. The instance c is the full image where the entity is masked out by replacing RGB values with zeros;
2. c is a small image patch randomly chosen around the entity. In practice, there are then several choices for c such that $c \in \mathcal{C}_e = \{c_1, c_2, ...\}$.

In both cases, the image patch c is processed by a ConvNet, parametrized by θ_1, to form an activation vector $u_c = \text{CNN}_{\theta_1}(c) \in \mathbb{R}^B$ (where B is the size of the last ConvNet filter, and equals 2048 in our experiments) obtained at the last layer of the network. The visual context vector $f_\theta(c) = Nu_c$ is then formed with the projection of u_c to the dimension d (of the textual space) with a matrix $N \in \mathbb{R}^{d \times B}$. Parameters to be learned are $\theta = \{\theta_1, N\}$.

3.3.1.3 Enhancing context with spatial information.

When a dataset provides localization information for entities (*i.e.* bounding boxes or segmentation masks), we can use these annotations as they provide additional spatial information. For example, by looking at the position of a cup in an image with respect to a table or the hand of a person, one can infer that cups lie on tables and that they can be handed by people. We consider two methods to model what we name *visual spatiality* to compute a vector $s_{(e,c)}$ representing the visual relationships between e and c. We then modify the function f_θ to integrate spatial information by defining a spatially-aware function f_θ^{sp} that integrate the spatial vector $s_{(e,c)}$ with a visual context element c as $f_\theta^{sp}(c, s_{(e,c)}) \in \mathbb{R}^d$. A summary of the options for modeling the spatial vector, and for the integration method, is presented in Table 3.1.

Representing spatial information

1. The first approach considers *low-level* features, and corresponds to a 4-d spatial vector whose components are the relative positions on the x and y

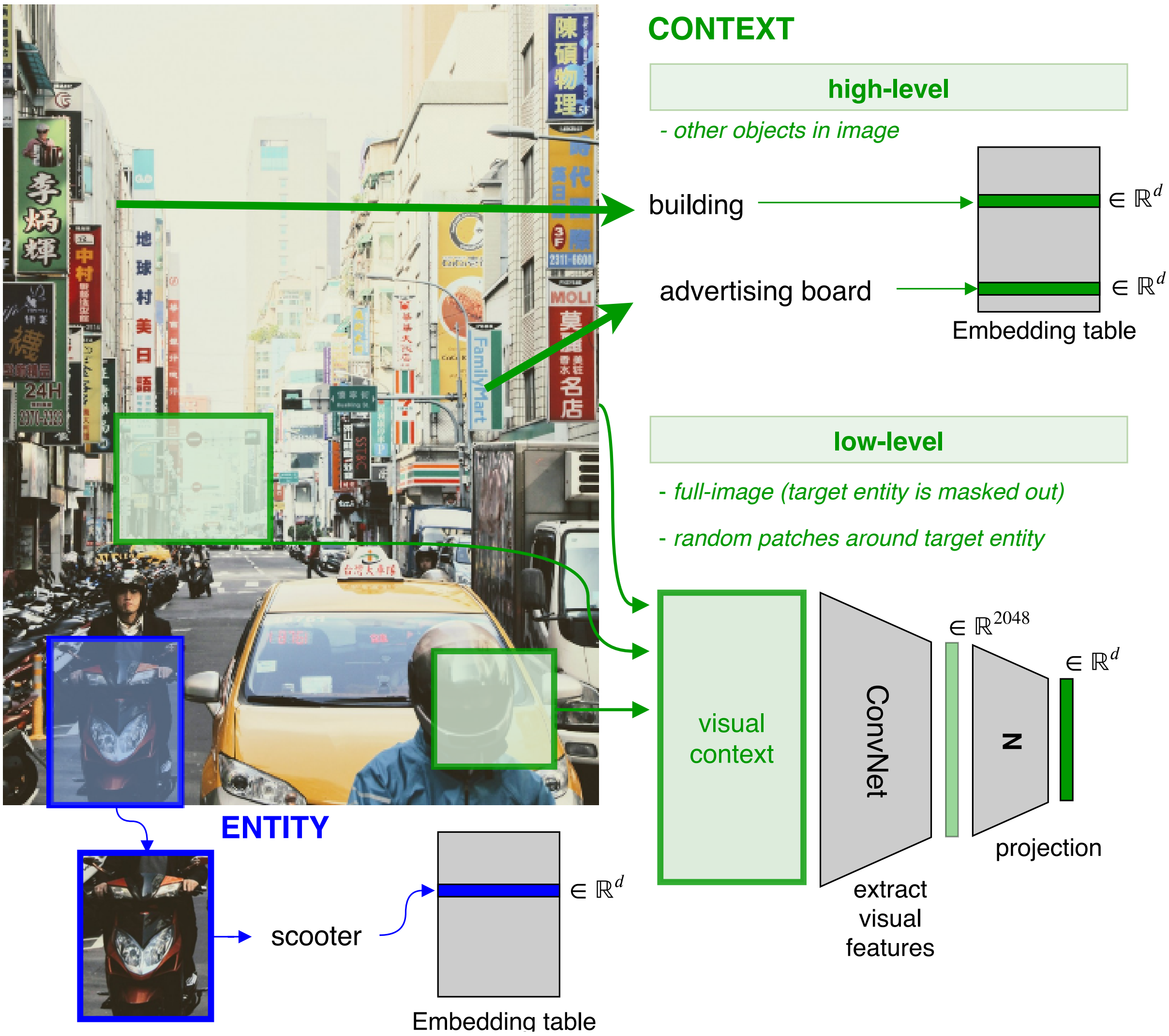


Figure 3.2 – **Overview of the model.** An object is chosen as the target entity and is represented with an embedding table. Two options are considered to define its context. (1) *high-level*: the list of the other objects in the image is known and the context is represented using an embedding table. (2) *low-level*: pixel values of surrounding image patches (or the full-image without the zone containing the entity) are processed by a Convolutional Neural Network (ConvNet) to compute a context vector.

axes of the two bounding boxes of the entity e and its context c (denoted δ_x and δ_y), as well as the ratio of width and height between the two bounding boxes of e and c (δ_{width} and δ_{height}).

2. Inspired by Ludwig et al. 2016, the second method builds a *high-level* features vector, corresponding to a 4-d spatial vector whose components are four indicator functions denoting whether the context c is below, beside, above, or bigger than the entity e (1 if true, 0 otherwise). The context is said to be “below" its entity if $|\delta_x| \leq \delta_y$, “above" its entity if $|\delta_x| \leq -\delta_y$ and “beside" otherwise. A context is said to be bigger than its entity if $\delta_{\text{width}}\delta_{\text{height}} \geq 1$.

We expect that the high-level modeling is sufficient to bring spatial semantics to the representations and that it will be easier to learn with. The low-level approach is expected to give finer-grain information, but on the other hand, it might bring noise during the learning phase.

Integrating spatial information Once the spatial vector $s_{(e,c)}$ is built, it is integrated with the visual context embedding $v_c = f_\theta(c)$, to form a spatially-informed visual context $v_c^{sp} = f_\theta^{sp}(c, s_{(e,c)})$ that is used in the `skip-gram` equations instead of $f_\theta(c)$. Again, two variants are considered, one simple *linear* combination and a *bilinear* one which might allow beneficial interactions to happen between textual and visual inputs:

1. A *linear* combination of the visual context v_c with the spatial vector $s_{(e,v)}$, *i.e.* $f_\theta^{sp}(c, s_{(e,c)}) = M.(v_c \oplus s_{(e,c)})$ where $M \in \mathbb{R}^{d\times(d+4)}$ and $\oplus$ denotes the concatenation operator;
2. A *bilinear* interaction $f_\theta^{sp}(c, s_{(e,c)}) = s_{(e,c)} M v_c$ where M is a 3-d tensor, *i.e.* $M \in \mathbb{R}^{4\times d\times d}$. This model has more free parameters but considers a bilinear interaction between the spatial vector $s_{(e,c)}$ and the visual context v_c.

$s_{(e,c)}$	Modeling spatial information in a 4-d vector 1. **Low-level** (L): $s_{(e,c)} = (\Delta_x, \Delta_y, \Delta_{\text{width}}, \Delta_{\text{height}})$ 2. **High-level** (H): $s_{(e,c)} = (\mathbb{1}_{\text{above}}, \mathbb{1}_{\text{below}}, \mathbb{1}_{\text{beside}}, \mathbb{1}_{\text{bigger}})$
f_θ^{sp}	Integration of the spatial vector with the context 1. **Linear** ($\oplus$): $f_\theta^{sp}(c, s_{(e,c)}) = M.(u_c \oplus s_{(e,c)})$, where $M \in \mathbb{R}^{d\times(d+4)}$ 2. **Bilinear** (b): $f_\theta^{sp}(c, s_{(e,c)}) = s_{(e,c)} M u_c$, where $M \in \mathbb{R}^{4\times d\times d}$

Table 3.1 – **Spatial information modeling and integration**. Two options are considered to represent spatial and size information, with a *high-level* or a *low-level* modeling. Besides, given a spatial 4d-vector, two options are considered to integrate spatial information with the context vector, a simple *linear* integration or a *bilinear* integration.

3.3.2 Integration in a multimodal model

We now present our multimodal representation learning model that integrates the previously presented visual module with the textual skip-gram . The main idea is that while word embeddings should be shared across modalities, context is media-specific. Indeed, imposing shared context representations would be overconstraining, because of the bias that affects contexts distribution across modalities, The contribution of each modality is controlled by a linear combination (hyperparameter α, determined by cross-validation) of modality-specific costs, which gives the following global loss function:

$$\mathcal{L}(T, U, \theta) = \mathcal{L}_t(T, U) + \alpha \mathcal{L}_i(T, \theta) \tag{3.2}$$

where T (resp. U) denotes the textual entity (resp. context) lookup table and $\mathcal{L}_t(T, U)$ is the Word2Vec loss function (Mikolov et al. 2013). $\mathcal{L}_i(T, \theta)$ is the visual skip-gram loss defined in Equation 3.1.

A crucial point is that this model does not require aligned texts and images to train the model, or extra pre-trained representations on external datasets – we only require that entities identified in images to be associated with a unique word of the vocabulary. Besides, we justify the use of a joint model as we think it is important that representations are learned both for entities and for contexts. Indeed, as the entities embeddings are affected by both modalities, the context representations should change and be updated by transitivity between modalities through the shared embeddings.

3.4 Evaluation protocol

In this section, we evaluate word embeddings on different tasks. In particular, we measure the performance of word embeddings built from visual data (RQ1) and multimodal data (RQ2).

3.4.1 Data

We use a large collection of English texts, a dump of the Wikipedia database (http://dumps.wikimedia.org/enwiki), cleaned and tokenized with Gensim (Rehurek et al. n.d.). This provides us with 4.2 million articles, and a vocabulary of 2.1 million unique words. As visual data, we use the Visual Genome dataset (Krishna et al. 2017) as it is a large image collection (108k images) with a large number of different objects (4842 unique entities with more than 10 occurrences) in rich and complex scenes (31 object instances per image on average).

3.4.2 Scenarios and Baselines

Scenarios and baselines are synthesized in Table 3.2. Specifically, we looked at the following model properties: (1) the kind of visual information being used (none, the visual appearance of the entities, or the visual context), (2) the specific modeling of visual context (high-level or low-level), (3) the way spatial information is handled (modeling and integration).

3.4.2.1 Scenarios

To evaluate the different components of our model, we evaluate different scenarios of the modelings proposed in Section 3.3.1.2. In particular, we train the model that uses other objects as visual contexts (*i.e.* the *high-level* context modeling, noted **O**), and models that use *low-level* contexts: either in the form of random image patches (noted **P**), either in the form of the full image (noted $\mathbf{P}_{\text{full}}$).

Models that use spatial context information are also evaluated and are denoted $\mathbf{Sp}(.,.,.)$ where the first argument denotes the visual context type (**O**, **P** or $\mathbf{P}_{\text{full}}$), the second the spatial context features (H for high-level, or L the low-level), and the third the integration method ($\oplus$ for concatenation and b for bilinear product). For instance, $\mathbf{Sp}(\mathbf{P}, L, b)$ corresponds to using image patches, with low-level spatial features and bilinear product.

All combinations of those models with the `skip-gram` text-only model (**T**) are trained and evaluated to get multimodal word representations, with the method explained in Section 3.3.2.

3.4.2.2 Baselines

Our baseline (**L**) is inspired by the state-of-the-art model of (Lazaridou et al. 2015), described in Section 2.2.2.1, since they use visual features from objects themselves to learn word representations in contrast to the visual context features we use in our model. For any visual entity e, they assume that a visual vector v_e representing the entity is available. During training, along with the purely-textual `skip-gram` loss, the similarity between the embedding t_e of the entity e and its visual appearance v_e is maximized with a max-margin loss:

$$\mathcal{L}_{\text{object}} = \sum_{e \in \mathcal{D}} \sum_{v^-} \max(0, \gamma - \cos(t_e, v_e) + \cos(t_e, v^-)) \tag{3.3}$$

where γ is the margin and v^- is the visual appearance of a "negative" object (randomly sampled over all objects, with uniform distribution). We note this model $\mathbf{L} + \mathbf{T}$ where **L** corresponds to the visual loss and **T** the text-only `skip-gram` loss.

To evaluate our visual context-driven multimodal representation learning model (RQ2), we also evaluate: 1) the `skip-gram` text only model (noted **T**), and 2) a sequential model, noted $\mathbf{O} \oplus \mathbf{T}$, where embeddings of model **T** are concatenated

	Text	**Visual appearance**	**Visual context**						
			Context modeling			**Spatial vector**		**Spatial integration**	
			Objects	Patches	Full image	Low-level	High-level	concat.	bilinear
L		✓							
O			✓						
$\mathbf{L} + \mathbf{O}$		✓	✓						
P				✓					
$\mathbf{P}_{\text{full}}$					✓				
$\mathbf{Sp}(\mathbf{O}, L, \oplus)$			✓			✓		✓	
$\mathbf{Sp}(\mathbf{O}, L, b)$			✓			✓			✓
$\mathbf{Sp}(\mathbf{O}, H, \oplus)$			✓				✓	✓	
$\mathbf{Sp}(\mathbf{O}, H, b)$			✓				✓		✓
T	✓								
$\mathbf{L} + \mathbf{T}$	✓	✓							
$\mathbf{O} \oplus \mathbf{T}$	✓		✓						
$\mathbf{O} + \mathbf{T}$	✓		✓						
$\mathbf{L} + \mathbf{O} + \mathbf{T}$	✓	✓	✓						
$\mathbf{P} + \mathbf{T}$	✓			✓					
$\mathbf{P}_{\text{full}} + \mathbf{T}$	✓				✓				
$\mathbf{Sp}(\mathbf{O}, L, \oplus) + \mathbf{T}$	✓		✓			✓		✓	
$\mathbf{Sp}(\mathbf{O}, L, b) + \mathbf{T}$	✓		✓			✓			✓
$\mathbf{Sp}(\mathbf{O}, H, \oplus) + \mathbf{T}$	✓		✓				✓	✓	
$\mathbf{Sp}(\mathbf{O}, H, b) + \mathbf{T}$	✓		✓				✓		✓

Table 3.2 – **Summary of scenarios and baselines**

with embeddings obtained from **O** and then projected in a lower-dimensional space with Principal Component Analysis (PCA). This serves as a comparison point between our joint approach and a sequential one.

3.4.3 Tasks

Similarly to previous work (Lazaridou et al. 2015; Collell et al. 2017), we evaluate our model on three different semantic tasks, namely word similarity and relatedness, feature norm prediction, and abstractness/concreteness prediction. Each task serves as a biased indicator of the quality of the embeddings. We refer the reader to Section 2.2.2.1 for a detailed description of the tasks.

3.4.4 Implementation details

Experiments use python and Tensorflow (Abadi et al. 2016). Images are upscaled to the shape 598×598 and passed through a pre-trained Inception-V3 ConvNet (Szegedy et al. 2016) to give spatial visual tensor of shape $17 \times 17 \times 2048$ (before the Rectified Linear Unit (ReLU) at the "Mixed_7c" layer). One slice of the tensor with a shape $1 \times 1 \times 2048$ corresponds to the activation of a region of the original image. We use 5 negative examples per entity, and our models are trained with Stochastic Gradient Descent (SGD) with learning rate $l_r = 10^{-3}$ and mini-batches of size 64. N and M are regularized with a L_2-penalty respectively weighted by scalars λ and μ. The values of hyperparameters were found with cross-validation: $\lambda = 0.1$, $\mu = 0.1$, $\gamma = 0.5$, $\alpha = 0.2$.

3.5 Experiments and Results

In Table 3.3 and Table 3.4, we report the results of the experiments for RQ1 discussing what kind of visual information can be useful. We analyze next the results to answer RQ2 and RQ3 in Table 3.5 and Table 3.6

RQ1: Evaluating visual context-driven semantic representations of words. Table 3.3 and Table 3.4 report results about the different visual information which can be used.

The first conclusion we draw is that surroundings of entities are more informative than the visual appearance of objects for word similarity benchmarks. Indeed, results of the word similarity task highlight that our model scenarios generally overpass baselines. For instance, results of our model $\mathbf{P}_{\text{full}}$ is on average 29% higher than those of the baseline **L**. However, on the feature-norm prediction task, direct visual features from objects (model **L**) are better suited for the categories

			VisSim	SemSim	Simlex	MEN	WordSim
Baseline		$\mathbf{L}$	43	45	16	22	17
Our models	Objects	$\mathbf{O}$	43	54	31	64	27
	Patches	$\mathbf{P}$	28	35	17	35	22
		$\mathbf{P}_{\text{full}}$	35	42	19	43	28
	Spatial	$\mathbf{Sp}(\mathbf{O}, L, \oplus)$	48	57	32	58	27
		$\mathbf{Sp}(\mathbf{O}, H, \oplus)$	48	58	30	58	25
		$\mathbf{Sp}(\mathbf{O}, L, b)$	46	56	**35**	54	28
		$\mathbf{Sp}(\mathbf{O}, H, b)$	**51**	**61**	33	62	30
	Ensemble	$\mathbf{L} + \mathbf{O}$	45	57	33	**66**	**34**

Table 3.3 – **RQ1 results — word similarity evaluation** Scores are Spearman correlations (multiplied by 100) on the word similarity benchmarks (only word pairs with visual entities are evaluated). Best results are highlighted in bold.

			Encyclopedic	Taste	Sound	Taxonomic	Function	Tactile	Color	Shape	Motion
Baseline		$\mathbf{L}$	56	49	36	**76**	**56**	**17**	**41**	**60**	**58**
Our models	Objects	$\mathbf{O}$	48	46	35	62	48	03	21	43	36
	Patches	$\mathbf{P}$	30	51	23	48	37	04	24	38	30
		$\mathbf{P}_{\text{full}}$	30	48	30	46	35	06	23	35	27
	Spatial	$\mathbf{Sp}(\mathbf{O}, L, \oplus)$	40	55	28	54	50	06	24	44	37
		$\mathbf{Sp}(\mathbf{O}, H, \oplus)$	40	**60**	33	54	50	11	25	41	34
		$\mathbf{Sp}(\mathbf{O}, L, b)$	37	57	27	50	50	15	24	38	32
		$\mathbf{Sp}(\mathbf{O}, H, b)$	38	58	27	58	47	10	22	43	34
	Ensemble	$\mathbf{L} + \mathbf{O}$	**58**	52	**42**	74	**56**	02	27	53	53

Table 3.4 – **RQ1 results — feature norm prediction** Scores are the f1-scores (multiplied by 100) at the feature-norm prediction task (grouped by feature category as proposed in (Collell et al. 2016)). Best results are highlighted in bold.

that describe visually the objects (*e.g. is_red* in 'Color' category or *is_round* in the 'Shape' category) but not for the other non visual categories such as 'Encyclopedic', 'Taste' and 'Sound'.

To measure the complementarity of the features from objects and from their surroundings, we also evaluated an ensemble model that combines the baseline **L** and the **O** model ($\mathbf{L} + \mathbf{O}$) where '+' denotes the summation of the loss functions when the embeddings are shared. Interestingly, combining visual contexts and di-

			VisSim	SemSim	Simlex	MEN	WordSim
Basel.	Text	$\mathbf{T}$	48	60	33	69	63
	Sequential	$\mathbf{O} \oplus \mathbf{T}$	49	62	33	71	64
	Joint	$\mathbf{L} + \mathbf{T}$	52	65	34	71	65
Our models	Objects	$\mathbf{O} + \mathbf{T}$	53	66	35	**75**	**67**
	Patches	$\mathbf{P} + \mathbf{T}$	53	65	35	72	**67**
		$\mathbf{P}_{\text{full}} + \mathbf{T}$	53	65	34	73	65
	Spatial	$\mathbf{Sp}(\mathbf{O}, L, \oplus) + \mathbf{T}$	52	66	36	73	64
		$\mathbf{Sp}(\mathbf{O}, H, \oplus) + \mathbf{T}$	54	66	35	72	64
		$\mathbf{Sp}(\mathbf{O}, L, b) + \mathbf{T}$	54	**68**	**38**	73	66
		$\mathbf{Sp}(\mathbf{O}, H, b) + \mathbf{T}$	**55**	67	34	**75**	64
	Ensemble	$\mathbf{L} + \mathbf{O} + \mathbf{T}$	54	66	35	**75**	65

Table 3.5 – **RQ2 experimental results — word similarity evaluation**. Scores are Spearman correlations (multiplied by 100) on the word similarity benchmarks

rect features ($\mathbf{L} + \mathbf{O}$) results in a model that has a very good average performance, showing the complementarity of visual contexts with visual entity representations.

Our second observation shows that using spatial information is useful: performance is better on the word similarity benchmarks, *i.e.* +9% improvement on average for $\mathbf{Sp}(\mathbf{O}, c, b)$ with respect to $\mathbf{O}$, and the feature-norm prediction task (+20%). Both high and low-level spatial features lead to similar results. This reinforces our intuition that visual context, and more particularly spatial information, are promising for learning word representation and reducing the *Human Reporting Bias* affecting texts and images.

The third conclusion we draw is that high-level contexts (in $\mathbf{O}$) yield better scores (+31%) than low-level contexts ($\mathbf{P}$ or $\mathbf{P}_{\text{full}}$). Using low-level visual features is a challenging problem. However, they are promising since they are cheap to collect, do not require context annotations, and contain rich information if handled correctly. The difficulty lies in the natural noise in the surroundings of objects and the need for visual modules that automatically extract high-level information from raw pixel values.

RQ2/RQ3: Evaluating our multimodal context-driven multimodal representation learning model / analysis. Table 3.5 and Table 3.6 reports the results to answer RQ2 and RQ3. Embeddings are initialized with pre-trained embeddings obtained from the text-only baseline.

Results highlight that all of the trained multimodal models outperform the text-only baseline on all evaluation tasks. For instance, $\mathbf{O} + \mathbf{T}$ shows an average improvement of 9% over $\mathbf{T}$. This is in-line with the conclusions of related works (Hill et al. 2014b). Besides, a joint model (*e.g.* $\mathbf{O} + \mathbf{T}$) compares favorably to

			Encyclopedic	Taste	Sound	Taxonomic	Function	Tactile	Color	Shape	Motion	Conc.
Basel.	Text	$\mathbf{T}$	58	52	44	79	62	11	32	54	60	42.1
	Sequential	$\mathbf{O} \oplus \mathbf{T}$	63	55	40	72	59	12	35	54	58	43.7
	Joint	$\mathbf{L} + \mathbf{T}$	61	55	42	80	59	11	31	54	62	43.4
Our models	Objects	$\mathbf{O} + \mathbf{T}$	62	55	46	**82**	61	13	33	55	61	42.9
	Patches	$\mathbf{P} + \mathbf{T}$	60	56	49	**82**	60	12	32	55	61	43.1
		$\mathbf{P}_{\text{full}} + \mathbf{T}$	60	55	44	**82**	**63**	14	32	55	59	43.2
	Spatial	$\mathbf{Sp}(\mathbf{O}, L, \oplus) + \mathbf{T}$	**64**	**59**	46	81	62	06	31	**57**	**63**	42.5
		$\mathbf{Sp}(\mathbf{O}, H, \oplus) + \mathbf{T}$	62	56	**52**	80	61	13	**34**	**57**	58	43.7
		$\mathbf{Sp}(\mathbf{O}, L, b) + \mathbf{T}$	63	56	48	81	60	13	32	56	**63**	42.5
		$\mathbf{Sp}(\mathbf{O}, H, b) + \mathbf{T}$	61	58	46	80	**63**	**15**	**34**	**57**	62	**44.4**
	Ensemble	$\mathbf{L} + \mathbf{O} + \mathbf{T}$	63	55	50	**82**	60	10	33	55	59	43.9

Table 3.6 – **RQ2 experimental results — feature norm and concreteness** Evaluation on the feature-norm and concreteness (conc.) prediction tasks. Scores for the feature-norm prediction task are f1-scores (multiplied by 100). Concreteness measures (conc.) are coefficients of determination (R^2) given in percentage.

a sequential model ($\mathbf{O} \oplus \mathbf{T}$) built from embeddings obtained from $\mathbf{O}$ and $\mathbf{T}$ as we note a 5% relative improvement, showing that embeddings computed using multiple modalities at once are beneficial. Like we did for RQ1, we also evaluated an ensemble model ($\mathbf{L} + \mathbf{O} + \mathbf{T}$) to measure the complementarity of visual features in the multimodal model. Again, we generally notice a slight improvement over both $\mathbf{O} + \mathbf{T}$ and $\mathbf{L} + \mathbf{T}$. This opens perspectives for formalizing and leveraging visual information from both entities and their context.

The obtained results are consistent with the conclusions drawn above on the RQ1 analysis: visual surroundings of entities are more useful than direct features on the evaluated tasks (3.2% improvement); the combination of both models shows the complementarity of the approaches, adding a spatial term for visual context significantly increases performances (6% improvement); finally, higher-level contexts are slightly easier to use than lower-level contexts (1% improvement).

Finally, to get a deeper insight into learned embeddings, we aim at explaining the impact of the visual modality on the multimodal word representation. To do so, with the model $\mathbf{O} + \mathbf{T}$, we estimate the correlation between the shift measured on the embedding (the norm of the difference of the initial textual embedding and the final multimodal embedding), and the concreteness degree of a word. We measured a correlation $\rho_{\text{Spearman}} = 0.33$, showing that visual and concrete words see their embeddings being more changed than other non visual and abstract words. This was to be expected because the visual part only adds information to visual entities.

3.6 Conclusion

3.6.1 Summary of the contributions

In this work, we proposed a multimodal (text and image) context-based approach to learn word embeddings. Through extensive experiments, and in line with related work, we observed the complementarity of visual and textual data to learn word representations. More importantly, we have shown that visual surroundings of objects and their relative localization are very informative to build word representations — actually, more than, but complementary to, the visual appearance of the objects themselves as exploited in previous works.

3.6.2 Perspectives

This work shows that visual information, in the form of visual contexts, can be integrated in a semantic space along with textual data. Extensions and future work could include the following perspectives.

Leverage other sources of information This chapter uses visual context to bring visual information in a semantic space. A natural extension is to consider additional complementary sources of information. For example, Knowledge Base (KB) contain curated common-sense knowledge about objects and their affordance. Exploiting a KB, along with visual contexts, could further improve multimodal word representations (Weston et al. 2013; Mancini et al. 2017).

Grounded relations While in this chapter we learn multimodal representations for words, a possible extension is to learn multimodal representations for relations, usually modeled with triplets *(subject, predicate, object)*. Several works exist to project triplets in a semantic space (A. Bordes et al. 2013; Toutanova et al. 2015) and some consider additional images to improve the quality of the learned space (Pezeshkpour et al. 2018). Using the visual context of relations could further improve the quality of the relation embeddings.

Comparing word representations There is no straightforward way to quantitatively evaluate word embeddings, and as a consequence several tasks and benchmarks have been proposed to measure the quality of semantics contained within word representations: similarity/relatedness benchmarks, analogy prediction task, feature norm prediction task. . . We empirically found that the various settings considered in this chapter (spatial information, high/low level modelings) produce embeddings that are relatively close. It appears difficult to find meaningful ways to quantitatively and qualitatively explore the difference between various word representation spaces, apart form separate evaluation on auxiliary tasks. Finding direct comparison methods remains an open-question.

CHAPTER 4

GROUNDING LANGUAGE IN THE VISUAL WORLD: THE CASE OF SENTENCES

Contents

Chapter abstract

In this chapter, we focus on learning grounded sentence representations. In related works, 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 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

- *sentences associated with the same visual content are close in the grounded space and*
- *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*.

4.1 Introduction

The previous chapter focuses on learning grounded word representations (Chapter 3). At another granularity level, having high-quality and general-purpose representations for sentences is crucial for many downstream applications, such as the ones used in machine translation (Bahdanau et al. 2015) or relation extraction (Wang et al. 2019) as detailed in Section 2.1.1.4. Moreover, as motivated in Section 2.2.2.2, 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).

From reviewing related works in Section 2.2.2.1 and Section 2.2.2.2, we observe that most approaches that leverage images to learn grounded word or sentence representations use cross-modal projections to incorporate visual semantics in the final representations (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 that the structure of the two spaces match. Indeed, Collell et al. 2018b empirically show that cross-modal projection of a source modality does not resemble the target modality in terms of neighborhood structure. This is especially the case for sentences and their associated images, where many different sentences can describe a similar image and vice-versa. Therefore, we argue that learning grounded representations with projections to a visual space is particularly inadequate in the case of sentences, as it might over-constraints the textual space and thus degrade learned representations. Regarding this issue, we formulate the following research question: how to incorporate visual semantics into sentence representations, without over-constraining the learned space?

To answer this question, we propose an 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.

- The *cluster information*: the implicit knowledge that sentences associated with the same image refer to the same underlying reality.

- The *perceptual information*, which is contained within high-level representations of images.

We aim at transferring, through the use of these two sources of information, 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.

In this chapter, we make the following contributions:

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 mean 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.

4.2 Incorporating visual semantics within an intermediate grounded space

4.2.1 Modeling motivation

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 imposed to be in one-to-one correspondence. Moreover, a caption can:

- (P1) have a wide variety of paraphrases and related sentences describing the same scene (e.g, *the kitten is devouring a mouse* vs. *a cat eating a mouse*),

- (P2) be visually ambiguous (e.g, *a cat is eating* can be associated with many different images, depending on the visual scene/context), or

- (P3) 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 (P3) through the projection function. They can handle (P1) by projecting related sentences to the same location in the visual space. However, they are over-sensitive to visual ambiguity (P2), because ambiguous sentences should be projected to different locations of the visual space, but this is not allowed by the model.

To overcome this lack of flexibility, we propose the following approach, illustrated in Figure 4.1. To cope with (P1), sentences associated with the same image should be close — we call this *cluster information*. To cope with (P2), given two pairs of sentence-image that are visually ambiguous, the distance between both sentences and between images should be similar since it measures the context discrepancy — but without requiring to project the text in a specific location. We call this *perceptual information*. Finally, as we want to preserve non-visual information in sentence representations (P3), 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.

4.2.2 Model overview

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, which balances a textual objective $\mathcal{L}_\mathcal{T}$ (taken from the existing literature) with an additional grounding objective $\mathcal{L}_\mathcal{G}$ (motivated in the previous section, and detailed in the next section):

$$\mathcal{L}(\theta^t, \theta^i) = \mathcal{L}_\mathcal{T}(\theta^t) + \mathcal{L}_\mathcal{G}(\theta^t, \theta^i) \tag{4.1}$$

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 sentence.

4.2.3 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

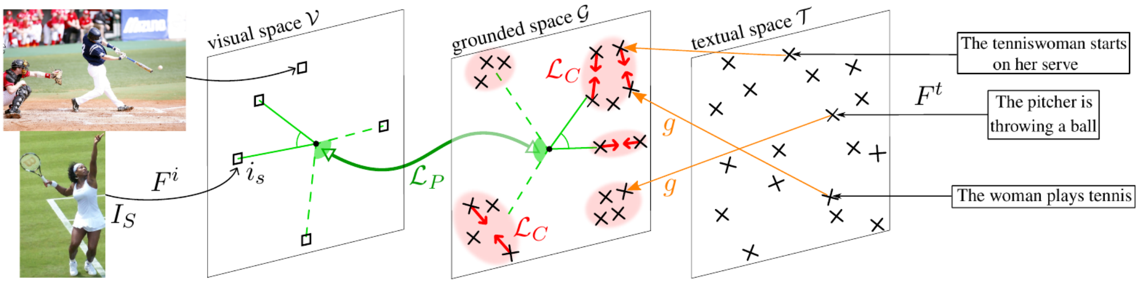

visual space $\mathcal{V}$
grounded space $\mathcal{G}$
textual space $\mathcal{T}$
The tenniswoman starts on her serve
The pitcher is throwing a ball
The woman plays tennis
F^t
F^i
i_s
I_S
$\mathcal{L}_P$
$\mathcal{L}_C$
$\mathcal{L}_C$
g
g

Figure 4.1 – **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 $\mathcal{L}_\mathcal{C}$ 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.

structure of the visual space. Thus, instead of directly applying the grounding objective 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 projected representation in the learned grounded space $g(s;\theta^i_g)$, noted $g(s)$ for simplicity, where g is a Multi-Layer Perceptron (MLP) with input $s = F^t(S;\theta^t)$ the sentence representation, and θ^i_g is its parameters ($\theta^i_g \subset \theta^i$).

We now describe the different grounding information (cluster and perceptual) and the corresponding losses composing the grounding objective $\mathcal{L}_\mathcal{G}$, applied in the grounded space.

Cluster information ($\mathbf{C}_g$) 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 (P1) 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 4.1, 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 between 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. 2015; 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 4.1) resulting in the cluster loss $\mathcal{L}_\mathcal{C}$:

$$\mathcal{L}_C = \sum_{(s,s^+,s^-)} \lfloor \gamma - \cos(g(s), g(s^+)) + \cos(g(s), g(s^-)) \rfloor_+ \tag{4.2}$$

where s^+ (resp. s^-) is a visually equivalent (resp. different) sentence to s. s^+ (resp. s^-) is randomly sampled with a uniform distribution over all visually equivalent (resp. different) sentences.

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 4.1 may be different, but they are both sports images, and thus their corresponding sentences should be somehow close in the grounded space. Defining a practical loss only supposes that we have a dataset of images associated with several captions, and we have many such datasets at our disposal.

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 4.1). 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 $\mathcal{L}_\mathcal{P}$:

$$\mathcal{L}_\mathcal{P} = -\rho(\{sim^{\text{text}}_{k_1,k_2}\}, \{sim^{\text{im}}_{k_1,k_2}\}) \tag{4.3}$$

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 The grounded space and cluster/perceptual information are combined into the grounding objective $\mathcal{L}_\mathcal{G}(\theta^t, \theta^i)$ as a linear combination of the aforementioned objectives:

$$\mathcal{L}_\mathcal{G}(\theta^t, \theta^i) = \alpha_P \mathcal{L}_P(\theta^t, \theta^i) + \alpha_C \mathcal{L}_C(\theta^t, \theta^i) \tag{4.4}$$

where α_P and α_C are hyper-parameters weighting contributions of $\mathcal{L}_P$ and $\mathcal{L}_C$. θ^i regroups weights of the image encoder F^i and weights θ^i_g of the projection function g.

4.3 Evaluation protocol

4.3.1 Datasets

Textual dataset. Following (Kiros et al. 2015; 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 Microsoft Common Objects in Context (MS COCO) (T. Lin et al. 2014) 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 amount of sentences in the training set of MS 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.

4.3.2 Baselines and Scenarios

In the experiments, we focus on one of the most established sentence models: `SkipThought` (noted $\mathbf{T}$), 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 $\mathbf{T}$, 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. 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) that embeds sentences in 1024 dimensional vectors.

Model Scenarios. We test variants of our grounding model presented in Section 4.2, all based on $\mathbf{T}$: $\mathbf{T}+\mathbf{C}_g$, $\mathbf{T}+\mathbf{P}_g$, $\mathbf{T}+\mathbf{C}_g+\mathbf{P}_g$, 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: $\mathbf{C}_g$, $\mathbf{C}_{id}$, $\mathbf{P}_g$, $\mathbf{P}_{id}$, $\mathbf{C}_g+\mathbf{P}_g$ and $\mathbf{C}_{id}+\mathbf{P}_{id}$.

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 Lazaridou et al. 2015 [1], this baseline learns to project sentences in the visual space using a max-margin loss:

[1] The original model is detailed in the Related Work chapter (Section 2.2.2.1)

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

where f is a MLP, γ' a fixed margin and i^- a non-matching image. Similarly to our scenarios, the sentence encoder is initialized with $\mathbf{T}$.
Sequential (**SEQ**): Inspired by Collell et al. 2017 [1], 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 sentence 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 Principal Component Analysis (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 (see Section 2.2.2.2): (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 Long-Short Term Memory (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.

4.3.3 Evaluation tasks and metrics

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

- **Semantic similarity benchmarks**: Semantic Textual Similarity (STS) and Sentences Involving Compositional Knowledge (SICK), as described in Section 2.2.2.2. Reported scores are spearman correlations between the cosine similarities of learned sentence embeddings and human-labeled scores.

- **Classification benchmarks**: Multi-Perspective Question Answering (MPQA), Movie Review (MR), SUBJ, Customer Reviews (CR) and Stanford Sentiment Treebank (SST), as described in Section 2.2.2.2. Reported scores are the classification accuracy for each of the learned evaluation classifier.

- **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 Collell et al. 2018b, that indicates the proportion of shared nearest neighbors between image representations and their corresponding captions in their respective spaces; mNNO measures the similarity between the neighborhood structures of two sets of paired vectors, *e.g.* mNNO $= 0.7$ means that corresponding textual and visual vectors share 70% of nearest neighbors in their respective space. To study *perceptual information*, we define ρ_{vis}, the Pearson correlation $\rho_{vis} = \rho(\cos(s, s'), \cos(v_s, v_{s'}))$ 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. All of the structural measures are computed either in the textual space for models with $g = id$ or in the grounded space when $g =$MLP.

4.3.4 Implementation details

Images are processed using a pre-trained `Inception-V3` network (Szegedy et al. 2016) ($d_i = 2048$). The model is trained with ADAM (Kingma et al. 2014) and a learning rate $l_r = 8.10^{-4}$. As done in (Kiros et al. 2015), our sentence encoder is a Gated Recurrent Unit (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 $d_t = 2048$.

4.4 Experiments and Results

Our main objective is to study the contribution brought by the visual modality to the grounded sentence representations, and 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 4.4.1), and quantitative results on the SentEval benchmark (Section 4.4.2).

4.4.1 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 4.1, we report the structural measures and the semantic relatedness scores of the baselines, namely **T** and **CM**,

		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 4.1 – **Intrinsic evaluations** carried out on the grounded space for models with $g = \text{MLP}$; 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)

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 in Table 4.1 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. This is promising as mNNO is considered as a realistic estimate of semantic similarity (Collell et al. 2018b): this is what we also verify experimentally, as $\mathbf{C}_g + \mathbf{P}_g$ shows improvement on the semantic relatedness tasks over the textual baseline **T**.

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.

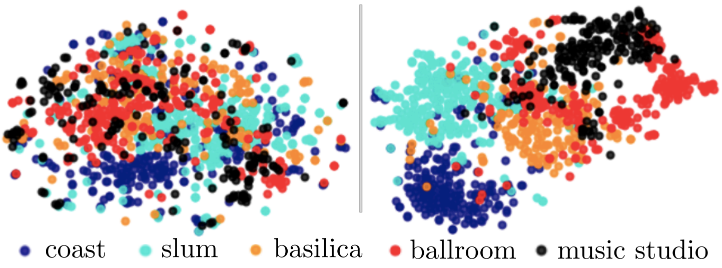


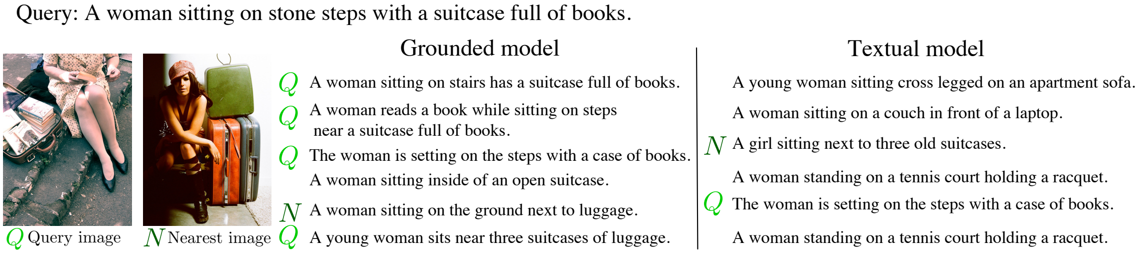
Figure 4.2 – t-distributed Stochastic Neighbor Embedding (t-SNE) visualization on Cross-Modal Places (CMPlaces) sentences for a set of randomly sampled visual scenes. Left: textual model **T**. Right: grounded model $\mathbf{C}_g + \mathbf{P}_g$.

To illustrate further, we performed a series of qualitative experiments that show that grounding groups together similar visual situations. Using sentences from 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 4.2 the corresponding sentences representations (in the grounded space) 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 4.2: $C_{inter} = 19, C_{intra} = 22$ for **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 **T**, which shows that clusters corresponding to different scenes are more clearly separated (resp. sentences corresponding to a given scene are more packed).

Furthermore, we show in Table 4.2 that concrete knowledge acquired via our grounded model can also be transferred to abstract sentences. To do so, we manually build abstract sentence queries 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 these examples, our grounded model is more accurate than the purely textual model to capture visual meaning. The observation that visual information

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 4.2 – **Qualitative sutdy.** Nearest neighbor of a given query among Flickr30K sentences.



Query: A woman sitting on stone steps with a suitcase full of books.

Grounded model

Q A woman sitting on stairs has a suitcase full of books.
Q A woman reads a book while sitting on steps near a suitcase full of books.
Q The woman is setting on the steps with a case of books.
A woman sitting inside of an open suitcase.
N A woman sitting on the ground next to luggage.
Q A young woman sits near three suitcases of luggage.

Textual model

A young woman sitting cross legged on an apartment sofa.
A woman sitting on a couch in front of a laptop.
N A girl sitting next to three old suitcases.
A woman standing on a tennis court holding a racquet.
Q The woman is setting on the steps with a case of books.
A woman standing on a tennis court holding a racquet.

Figure 4.3 – **Qualitative study.** Nearest neighbors of a selected sentence in the validation set of MS COCO, for both grounded and purely textual models. Q is the image corresponding to the query, N is the nearest neighbor of Q in the visual space.

propagates from concrete sentences to abstract ones is analogous to findings made in previous research on word embeddings (Hill et al. 2014a).

Finally, to illustrate the discrepancy on the mNNO metric observed between $\mathbf{C}_g + \mathbf{P}_g$ and **T**, we select a query image Q in the validation set of MS COCO, along with its corresponding caption S; we display, in Figure 4.3, 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 of S is mostly made of sentences corresponding to Q or N.

Hypotheses validation We now validate our hypotheses that motivated the modeling of the grounded space (cf. Section 4.2.1), using the **CM** baseline and our model scenarios as outlined in Table 4.1. 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 evaluations respectively correspond to the rows **CM** (text) and **CM** (vis.) in Table 4.1. Results highlight that:

1. Using a grounded space is beneficial; indeed, semantic relatedness and mNNO scores are higher in the lower half of Table 4.1, 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

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†	**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‡	**T + 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 4.3 – **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.

conjunction of both perceptual and cluster information lead 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.

4.4.2 Downstream evaluation: transfer tasks

We now turn to evaluate embeddings on extrinsic tasks. Table 4.3 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 except MRPC, 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 (like in the Chapter 3). 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.

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 ($g = id$). We believe that such an approach is key to integrate information from different modalities.

4. Among our model scenarios, $\mathbf{T} + \mathbf{P}_g$ has maximal scores on the most tasks; however, it shows lower scores on Stanford Natural Language Inference (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. Finally, the combined model $\mathbf{T} + \mathbf{C}_g + \mathbf{P}_g$ shows a good balance between classification and entailment tasks.

4.5 Conclusion

4.5.1 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) the definition and the use of both perceptual and cluster information, and (2) the modeling of an intermediate grounded space enabling to relax the constraints on the textual space. Moreover, we validate our hypotheses with quantitative and qualitative results against purely textual baselines on a variety of natural language tasks.

4.5.2 Perspectives

As future work, we plan to use visual information to specifically target complex downstream tasks requiring common-sense and reasoning such as the question answering or visual dialogue tasks.

Moreover, we could investigate the use of videos instead of images because of their temporal aspect (since sentences often describe actions grounded in time) and because multiple frames may bring a better visual context than a single image.

Finally, we would like to gain a deeper understanding of the use of a pearson correlation as a training objective. This training objective could potentially be used in other applications.

Chapter 5

LEVERAGING LANGUAGE FOR VISUAL UNDERSTANDING

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 and challenging approach, context-aware zero-shot learning, 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 approach 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*.

5.1 Introduction

In this chapter, we tackle the second axis of the thesis and we bring elements of response to the research question: *can language help visual recognition?*. As discussed in Chapter 2, and in particular in Section 2.3.2, linguistic representations can be used to augment capacities of computer vision recognition systems, typically when visual supervision is scarce. We thus focus on the zero-shot recognition task, where no visual supervision is available for a set of classes, and we explore how linguistic representations can help for this task (see Figure 5.1).

5.1.1 Zero-shot recognition

Traditional Computer Vision models, such as Convolutional Neural Network (ConvNet) (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. 2017; Real et al. 2018) 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. 2009; Mensink et al. 2012; Y. Fu et al. 2015; Kodirov et al. 2017). In the object recognition field, ZSL aims at labeling an instance of a 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.

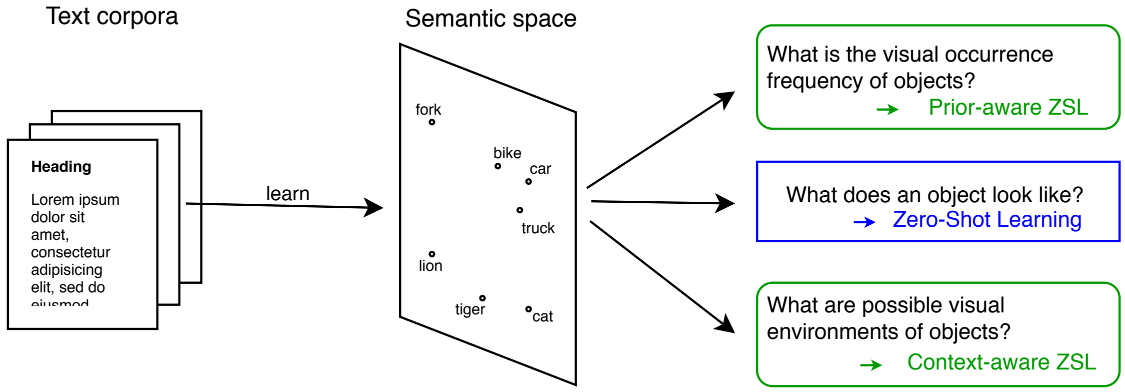

Text corpora
Heading
Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod
learn
Semantic space
fork
bike
car
truck
lion
tiger
cat
What is the visual occurrence frequency of objects?
→ Prior-aware ZSL
What does an object look like?
→ Zero-Shot Learning
What are possible visual environments of objects?
→ Context-aware ZSL

Figure 5.1 – **Intuitions of the context-aware ZSL approach** from a language perspective. Given a linguistic corpus, and a subsequently learned semantic space representing concepts, one can ask several questions regarding visual information of these concepts. The blue box contains the Research Question (RQ) asked in the traditional ZSL setting and we additionally consider two other RQ (in green boxes) that we attempt to answer in this chapter.

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

5.1.2 Leveraging visual context for ZSL

The intuitive principle that *some objects tend to be found in some contexts but not others*, is at the core of many works. Some works in Computer Vision have exploited visual context to refine the predictions of classification (Mensink et al. 2014), detection (Bell et al. 2016), or segmentation (F. Yu et al. 2016) models. Visual context can either be *low-level* (*i.e.* raw image pixels) or *high-level* (*i.e.* labeled objects):

- When visual context is exploited in the form of *low-level* information (Torralba 2003; Wolf et al. 2006; Torralba et al. 2010), it often consists of global image features. For instance, in (X. He et al. 2004), a Conditional Random Field (CRF) is trained at combining low-level image features to assign to each pixel a class.
- In *high-level* approaches, the referential meaning of the context objects (*i.e.* class labels) is used. For example, Rabinovich et al. 2007 show that high-level context can be used at the post-processing level to reduce the ambiguities of a pre-learned object classification model, by leveraging co-occurrence patterns between objects that are computed from the training set.

We believe that visual context, *i.e.* the other entities surrounding an object, also explains human's ability to recognize an object that they have never 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. To the best of our knowledge, context has not been exploited in ZSL because 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 chapter, our goal is to leverage visual context as an additional source of knowledge for ZSL, by exploiting the distributed word representations (Mikolov et al. 2013) of the object class labels. Overall, we formulate, and strive to answer, the following RQ (illustrated in Figure 5.1).

RQ1: Is information about **possible visual environment** of objects contained in language in general, and in semantic representation in particular?

To tackle this question, we adopt a probabilistic framework in which the probability to recognize a given object is split into three components:

1. a *visual component* based on its visual appearance (which can be derived from any traditional ZSL approach),
2. a *contextual component* exploiting its visual context,
3. a *prior component*, which estimates the frequency of objects in the dataset.

As a complementary contribution, as all classes of the target domain are not equally likely (non-uniform prior), we ask the following RQ (illustrated in Figure 5.1):

RQ2: Is information about **visual occurrence frequency** of objects contained in language in general, and in semantic representation in particular?

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.

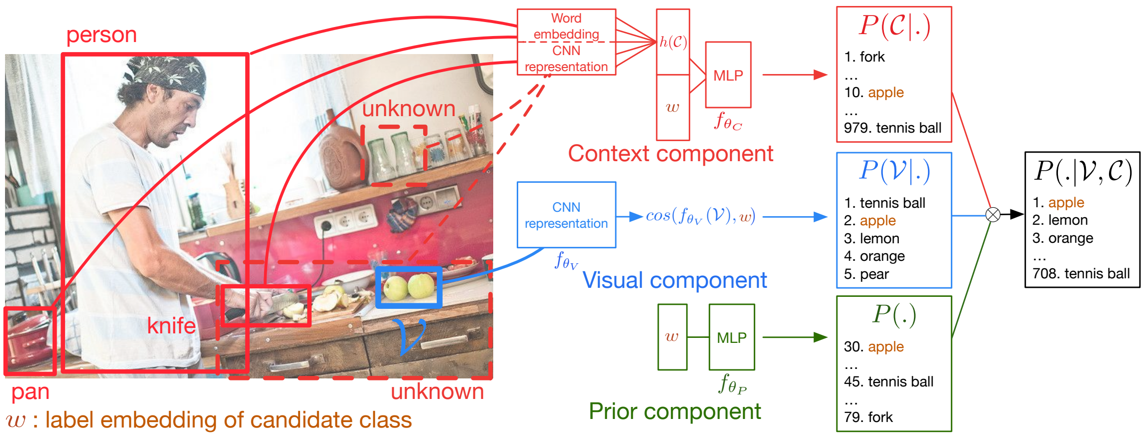


Figure 5.2 – **Model Overview.** The goal is to find the class (in the target domain) of the object contained within the blue image region $\mathcal{V}$. Its context is formed of labeled objects from the source domain (red plain boxes) and of unlabeled object from the target domain (red dashed boxes). These objects can either be represented with word embeddings (when they are known), or ConvNet representations.

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.

5.2 Context-aware Zero-Shot Learning

5.2.1 Model overview

Intuition The intuition behind our approach is illustrated in Figure 5.2, where the blue box contains the object of interest. Here, the class is *apple*, which belongs to the target domain $\mathcal{T}$. Three independent component come to play:

- 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.

- the context component, which is discriminant in that case: 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.

Notations and definitions 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 zero-shot learning* — 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 $\mathcal{V}$, 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.

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 5.4) that this hypothesis is acceptable. This hypothesis leads to the following expression:

$$P(i|\mathcal{V}, \mathcal{C}) \propto P(\mathcal{V}|i)P(\mathcal{C}|i)P(i) \tag{5.1}$$

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.

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. 2013), 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 5.2.3). This allows dealing with imbalanced datasets, in contrast to ZSL models like `DeViSE` (Frome et al. 2013). This intuition is experimentally validated in Section 5.4.2.

5.2.2 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 5.1. 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 5.2.3. Finally, we explain how we combine them to produce the global probability $P(.|\mathcal{C}, \mathcal{V})$ in Section 5.2.4.

5.2.2.1 Visual component

The visual component models $P(\mathcal{V}|i)$, *i.e.* 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: $f_{\theta_V}(\mathcal{V}) = W_V.\texttt{CNN}(\mathcal{V}) + b_V \in \mathbb{R}^d$ where $\texttt{CNN}(\mathcal{V})$ is a vector in $\mathbb{R}^{d_{\text{visual}}}$, output by a pre-trained ConvNet 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 of i:

$$\log P(\mathcal{V}|i;\theta_V) \propto \cos(f_{\theta_V}(\mathcal{V}), w_i) := \log \widetilde{P}_{visual} \tag{5.2}$$

5.2.2.2 Context component

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

$$\begin{aligned} \log P(\mathcal{C}|i;\theta_C) \propto f_{\theta_C}(\mathcal{C}, w_i) &= f_{\theta_C^1}(h_{\theta_C^2}(\mathcal{C}) \oplus w_i) \\ &:= \log \widetilde{P}_{context} \end{aligned} \tag{5.3}$$

where $h_{\theta_C^2}(\mathcal{C}) \in \mathbb{R}^d$ is a vector representing the context (detailed in the paragraph below), $\theta_C = \{\theta_C^1; \theta_C^2\}$ 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 Multi-Layer Perceptron (MLP). We found that concatenating $h_{\theta_C^2}(\mathcal{C})$ with w_i leads to better results than using a cosine similarity, as it is done for the visual component (in Equation 5.2).

Modeling of $h_{\theta_C^2}(\mathcal{C})$ To specify the modeling of $h_{\theta_C^2}(\mathcal{C})$, we propose various *context models* depending on which context objects are considered (source or target domain) and how they are represented (high/low-level representation). An illustration of the possible choices is depicted in Figure 5.3. 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:

1. The *low-level* (L) approach takes into account the representation from the image region $\mathcal{V}_k$ of a context object. That context object can belong to either the source or the target domain, and we further distinguish these cases:

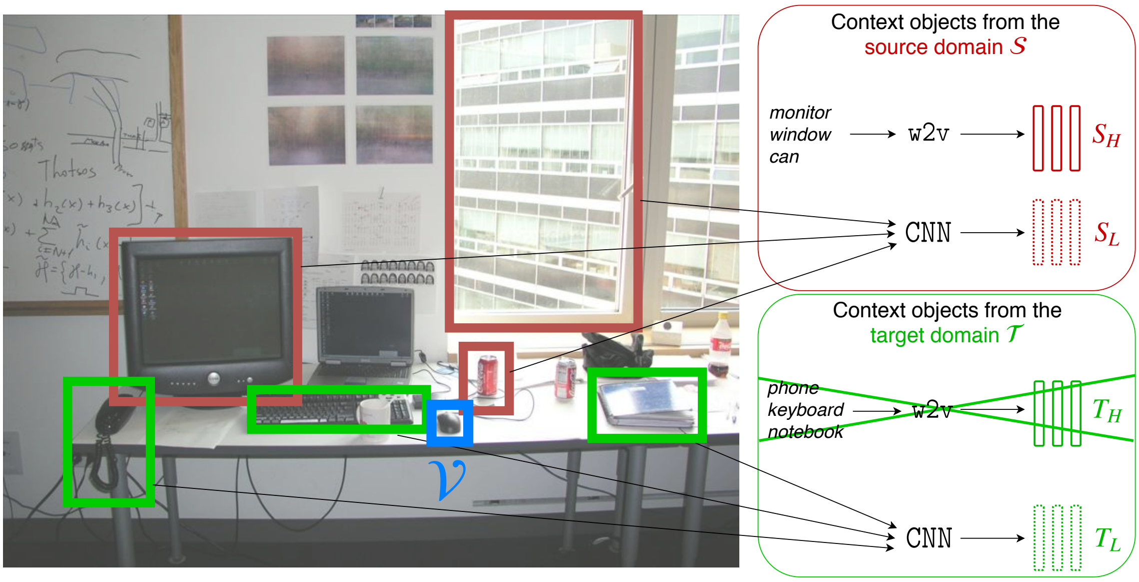


Figure 5.3 – **Presentation of context models.** The object of interest is in the blue box $\mathcal{V}$, which contains a *computer mouse* in this case. The classes of the context objects, for which we assume to have bounding boxes, can either belong to the *source domain* $\mathcal{S}$ or the *target domain* $\mathcal{T}$. For objects of the source domain, in red boxes, we can use the semantic representation of their label or use the visual representation of the content of the bounding box, thanks to a pre-trained ConvNet. For objects of the target domain, in green boxes, since labels are unknown, only the visual representation of the content of the box can be used.

- the context object belongs to the *source domain* $\mathcal{S}$, this produces the context model S_L:

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

- or the context object belongs to the *target domain* $\mathcal{T}$, this produces the context model T_L:

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

2. 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). Again, depending on the domain in which the context object belongs, we distinguish:

- the context object belongs to the *source domain* $\mathcal{S}$, this produces the context model S_H:

$$S_H = \{w_k | k \in \mathcal{C} \cap \mathcal{S}\}$$

- or the context object belongs to the *target domain* $\mathcal{T}$, this produces the context model T_H:
$$T_H = \{w_k | k \in \mathcal{C} \cap \mathcal{T}\}$$

  Note that T_H is not defined in the zero-shot setting, since class labels of objects from the target domain are unknown; we only use it to define Oracle models (Section 5.3.3).

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, $h_{\theta_C^2}$ averages the representations of these vectors to build a global context representation. For example:

$$h_{\theta_C^2}(\mathcal{C}_{S_H \cup T_L}) = \frac{1}{|\mathcal{C}_{\mathcal{S}}| + |\mathcal{C}_{\mathcal{T}}|}\Big[\underbrace{\sum_{(i,\mathcal{V}_i)\in\mathcal{C}_{\mathcal{S}}} w_i}_{S_H} + \underbrace{\sum_{(j,\mathcal{V}_j)\in\mathcal{C}_{\mathcal{T}}} (W_C.\texttt{CNN}(\mathcal{V}_j) + b_C)}_{T_L}\Big]$$

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

5.2.2.3 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:

$$\log P(i; \theta_P) \propto f_{\theta_P}(w_i) := \log \widetilde{P}_{prior} \tag{5.4}$$

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

Note that works have shown that there was a relationship between the term frequency (in text) and their representation (Schakel et al. 2015a). We here investigate whether textual representations contain information about visual occurrence frequencies.

5.2.3 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 5.4.2). 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:

5.2.3.1 Prior objective ($\mathcal{L}_P$)

For this objective, negative object classes are sampled from the *uniform* distribution U:

$$\mathcal{L}_P = \mathop{\mathbb{E}}_{i\sim P^\star}\mathop{\mathbb{E}}_{j\sim U}\left\lfloor \gamma_P - f_{\theta_P}(w_i) + f_{\theta_P}(w_j)\right\rfloor_+ \tag{5.5}$$

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.

5.2.3.2 Visual and context objectives ($\mathcal{L}_V$ and $\mathcal{L}_C$)

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

$$\mathcal{L}_V = \mathop{\mathbb{E}}_{i,\mathcal{V}\sim P^\star}\mathop{\mathbb{E}}_{j\sim P^\star}\left\lfloor \gamma_V - f_{\theta_V}(\mathcal{V})^\top w_i + f_{\theta_V}(\mathcal{V})^\top w_j\right\rfloor_+ \tag{5.6}$$

$$\mathcal{L}_C = \mathop{\mathbb{E}}_{i,\mathcal{C}\sim P^\star}\mathop{\mathbb{E}}_{j\sim P^\star}\left\lfloor \gamma_C - f_{\theta_C}(\mathcal{C}, w_i) + f_{\theta_C}(\mathcal{C}, w_j)\right\rfloor_+ \tag{5.7}$$

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

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

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 5.4.2), hence improving the `DeViSE` model itself by a proper modeling of the different probabilities at hand.

5.2.4 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 5.2.3, 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 computed respectively. However, the normalization constants in Equations 5.2, 5.3 and 5.4, 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:

$$P(.|\mathcal{V}, \mathcal{C}) = \underbrace{(\widetilde{P}_{context})^{\alpha_C}}_{P(\mathcal{C}|.)} . \underbrace{(\widetilde{P}_{visual})^{\alpha_V}}_{P(\mathcal{V}|.)} . \underbrace{(\widetilde{P}_{prior})^{\alpha_P}}_{P(.)} \tag{5.8}$$

To see whether this hypothesis is reasonable, we did some *post-hoc* analysis of one of our model, and plotted in Figure 5.4 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, $P(.|\mathcal{C}) = (\widetilde{P}_{context})^{\alpha_C}(\widetilde{P}_{prior})^{\alpha_P}$.

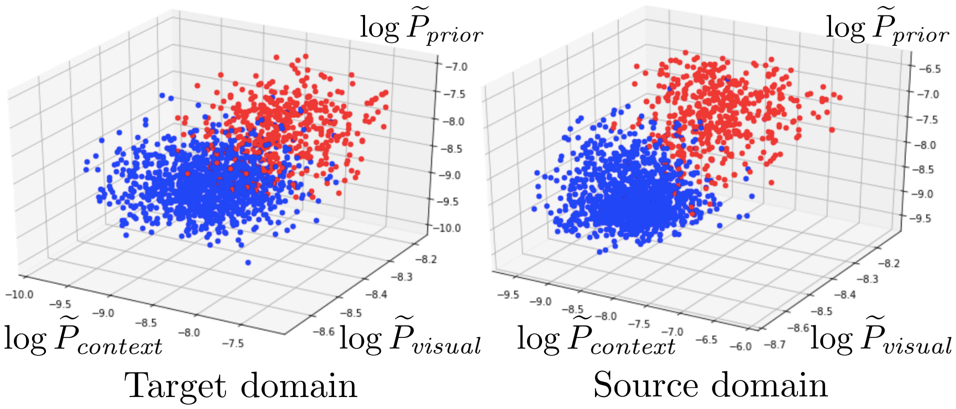


Figure 5.4 – **3D visualization of the unnormalized log-probabilities** for each component ($N = 500$). Context model $S_L \cup S_H \cup T_L$. Correct configurations are represented as red points and randomly sampled incorrect configurations are represented as blue points.

5.3 Experimental protocol

5.3.1 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. 2009), SUN (Xiao et al. 2010), CUB-200 (Wah et al. 2011) or LAD (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

	Raw dataset	Adapted dataset
Number of images	108K	108K
Number of unique entities	105K	4842
Number of object instances	3.8M	3.4M
Number of instances per image (average)	35	31

Table 5.1 – **Dataset statistics** Left: original statistics (Krishna et al. 2017); Right: statistics of our adapted dataset

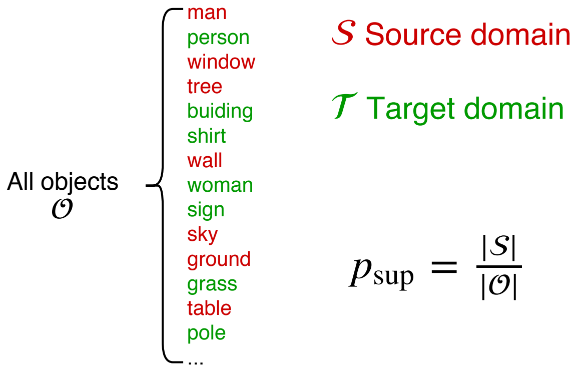

All objects
$\mathcal{O}$
man
person
window
tree
buiding
shirt
wall
woman
sign
sky
ground
grass
table
pole
...
$\mathcal{S}$ Source domain
$\mathcal{T}$ Target domain
$p_{\text{sup}} = \frac{|\mathcal{S}|}{|\mathcal{O}|}$

Figure 5.5 – **Randomly splitting objects in source and target domains**

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. See the dataset statistics in Table 5.1.

5.3.2 Evaluation methodology and metrics

We define the *level of supervision* p_{sup} as the ratio of the size of the source domain over the total number of objects: $p_{\text{sup}} = |\mathcal{S}|/|\mathcal{O}|$. For a given p_{sup} ratio, the source $\mathcal{S}$ and target $\mathcal{T}$ domains are built by randomly splitting $\mathcal{O}$ accordingly (see Figure 5.5). Every object is annotated with a bounding box and we use this supervision in our model for entities of both source and target domains.

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). In addition, we also report scores in the *generalized* ZSL setting, where entities are retrieved among all entities (from both the source and target domains).

The model's prediction takes the form of a list of n classes, sorted by probability (as seen in Figure 5.6) ; 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 $\text{FR} = \frac{2}{n-1}(r-1)$. To further evaluate the performance over the whole test set, the MFR metric is used (Fuhr 2017). It is computed by taking the mean value of 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 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

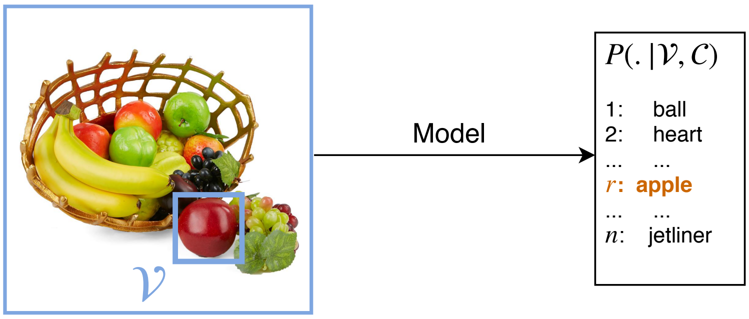


Figure 5.6 – **Illustration of the First Relevant (FR) metric.** r is the rank of the correct class, $r = 1$ if the model returns the correct class and $r = 50\%$ if the model is random. The Mean First Relevant (MFR) metric we used is a linear rescaling of the FR metric, averaged over the whole test set.

for meaningful comparison with a varying p_{sup}. For the sake of completeness, we also measure MRR and Recall@k metrics.

5.3.3 Scenarios and Baselines

5.3.3.1 Model scenarios

Model scenarios depend on the information that is used in the probabilistic setting: $\varnothing$ (prior only), $\mathcal{C}$ (prior + context), $\mathcal{V}$ (prior + visual area) or both $\mathcal{C}$ and $\mathcal{V}$ (prior + visual area + visual context). When contextual information is involved, a context model $\star$ is specified to represent $\mathcal{C}$, which we note $\mathcal{C}_\star$. As explained in Section 5.2.2.2, the different context models are any combination of S_H, S_L and T_L:

$$\star \in \{S_H, S_L, T_L, S_L \cup T_L, S_H \cup T_L, S_L \cup S_H \cup T_L\}$$

For clarity's sake, we note the various model scenarios with the letter M. For example, $\mathrm{M}(\mathcal{C}_{S_H \cup T_L}, \mathcal{V})$ models the probability $P(\mathcal{C}_{SH} \cup T_L|.)P(\mathcal{V}|.)P(.)$ as explained in Section 5.2.4, and $\mathrm{M}(\mathcal{V})$ models $P(\mathcal{V}|.)P(.)$, and $\mathrm{M}(\varnothing)$ models $P(.)$.

5.3.3.2 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.

- *True Prior:* This Oracle uses, for its prior component, the true prior distribution $P^\star(i) = \frac{\#i}{M}$ 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. This oracle measures the performance of a naive model, which ignores visual context, and which would have access to the true distribution for objects of both the source and target domains.

- *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 $P_{\text{co-oc}}(c|i) = \frac{\#(c,i)M}{\#c\#i}$ is the probability that objects c and i co-occur in images, with $\#(c,i)$ the number of co-occurrences of c and i. This oracle measures the performance of a naive-bayes model which would have access to the true object distribution and co-occurrence statistics for objects of both the source and target domains. While the modeling of the contextual information is naive as it only considers pairwise interaction (assuming conditional independences), we expect that this oracle would give a raw upper-bound estimate of the benefit brought by leveraging visual context.

- *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. This oracle naively uses purely-textual occurrence and co-occurrence statistics, in the prior and context component, for objects of both the source and target domains. Despite the use of co-occurrence knowledge for objects of the target domain, we expect to show that textual statistics are not sufficient to satisfyingly leverage visual contextual information. This would further illustrate the text-image reporting bias detailed in Section 2.2.1.

- *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. In the zero-shot setting, context objects of the target domain $\mathcal{T}$ are not labeled by definition but here we assume that their labels are known and we can then use T_H. This oracle measures performances that could be reached if the visual context for an object was perfectly known, by assuming that surroundings objects of the target domain are known.

5.3.3.3 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 5.6), $f_{\theta_V}(\mathcal{V})$ by the concatenation ($\oplus$) of $h(\mathcal{C})$ and $f_{\theta_V}(\mathcal{V})$ projected in $\mathbb{R}^d$ with a 2-layer MLP.

- DeViSE$(\mathcal{V})$: To evaluate the impact of our Bayesian model (Equation 5.1) and our sampling strategy (Section 5.2.3), we compare against DeViSE (Frome et al. 2013). DeViSE$(\mathcal{V})$ is different from M$(\mathcal{V})$ because negative examples in $\mathcal{L}_V$ are uniformly sampled, and the prior $P(.)$ is not learned. DeViSE directly models $P(. \mid \mathcal{V})$ while M$(\mathcal{V})$ models $P(\mathcal{V} \mid .)P(.)$.

- DeViSE$(\mathcal{C} \oplus \mathcal{V})$: similarly to M$(\mathcal{C} \oplus \mathcal{V})$, we define a baseline that does not rely on the conditional independence of $\mathcal{C}$ and $\mathcal{V}$. To do so, we replace, in the expression of $\mathcal{L}_V$ (Equation 5.6), $f_{\theta_V}(\mathcal{V})$ by the concatenation ($\oplus$) of $h(\mathcal{C})$ and $f_{\theta_V}(\mathcal{V})$ projected in $\mathbb{R}^d$ with a 2-layer MLP. Unlike in the M$(\mathcal{C} \oplus \mathcal{V})$ oracle, we use the same sampling strategy as DeViSE .

- 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.

5.3.4 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. 2013) on Wikipedia. Image regions are cropped, rescaled to (299×299), and fed to CNN, an Inception-V3 ConvNet (Szegedy et al. 2016), whose weights are kept fixed during training. This model is pre-trained on ImageNet (Deng et al. 2009). 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. 2014) 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.

5.4 Results

5.4.1 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 ($\star = S_H$) and

		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*
$\mathrm{M}(\varnothing)$	$P(.)$	38.6	23.7	13.8	12.0	10.6	11.2
$\mathrm{M}(\mathcal{V})$	$P(\mathcal{V}\vert.)P(.)$	20.5	10.7	6.0	1.5	2.6	3.6
$\mathrm{M}(\mathcal{C}_{S_H})$	$P(\mathcal{C}\vert.)P(.)$	28.7	14.4	9.1	4.2	4.3	4.4
$\mathrm{M}(\mathcal{C}_{S_H}, \mathcal{V})$	$P(\mathcal{C}\vert.)P(\mathcal{V}\vert.)P(.)$	**18.1**	**9.0**	**5.2**	**1.1**	**1.9**	**2.4**
δ_C (%)		*11.6*	*16.4*	*12.1*	*23.7*	*27.3*	*31.5*

Table 5.2 – **Evaluation of various information sources, with varying levels of supervision.** MFR scores in %. δ_C is the relative improvement (in %) of $\mathrm{M}(\mathcal{C}_{S_H}, \mathcal{V})$ over $\mathrm{M}(\mathcal{V})$. Entities are retrieved only among entities of the domain at hand.

report MFR results with $p_{\text{sup}} = 10, 50, 90\%$ in Table 5.2 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. We also report results on the MRR and top-k metrics for the same models in Table 5.4 and the results in the generalized setting in Table 5.3.

Results highlight that contextual knowledge acquired from the source domain can be transferred to the target domain, as $\mathrm{M}(\mathcal{C}_{S_H})$ significantly outperforms the *Random* baseline. As expected, it is not as useful as visual information: $\mathrm{M}(\mathcal{V}) \overset{\text{MFR}}{<} \mathrm{M}(\mathcal{C}_{S_H})$, where $\overset{\text{MFR}}{<}$ means lower MFR scores, *i.e.* better performances. However, Table 5.2 demonstrates that contextual and visual information are complementary: using $\mathrm{M}(\mathcal{C}_{S_H}, \mathcal{V})$ outperforms both $\mathrm{M}(\mathcal{C}_{S_H})$ and $\mathrm{M}(\mathcal{V})$ (for example with $p_{\text{sup}} = 50\%$, $\mathrm{M}(\mathcal{C}_{S_H}, \mathcal{V})$ reaches 9.0% MFR while $\mathrm{M}(\mathcal{C}_{S_H})$ reaches 14.4% MFR and $\mathrm{M}(\mathcal{QV})$ 10.7% MFR). Interestingly, as the learned prior model $\mathrm{M}(\varnothing)$ is also able to generalize, we show that visual frequency can somehow be learned from textual semantics, which extends previous work where word embeddings were shown to be a good predictor of textual frequency (Schakel et al. 2015b).

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 $\mathrm{M}(\mathcal{C}_{S_H})$ on $\mathcal{T}$
- the increasing relative improvement δ_C of $\mathrm{M}(\mathcal{C}_{S_H}, \mathcal{V})$ over $\mathrm{M}(\mathcal{V})$ 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

		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*
$\mathrm{M}(\varnothing)$	$P(.)$	39.6	26.3	16.9	6.6	8.68	10.9
$\mathrm{M}(\mathcal{V})$	$P(\mathcal{V}\|.)P(.)$	21.0	11.8	6.9	0.9	2.3	3.5
$\mathrm{M}(\mathcal{C}_{S_H})$	$P(\mathcal{C}\|.)P(.)$	28.6	15.0	10.7	3.5	3.9	4.4
$\mathrm{M}(\mathcal{C}_{S_H}, \mathcal{V})$	$P(\mathcal{C}\|.)P(\mathcal{V}\|.)P(.)$	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 5.3 – **MFR scores in the generalized ZSL setting.** Entities are retrieved among every possible entities (from both the source and target domain)

	Target domain $\mathcal{T}$				Source domain $\mathcal{S}$			
	top-1	top-5	top-10	MRR	top-1	top-5	top-10	MRR
Random	<0.1	0.2	0.4	<0.1	<0.1	0.2	0.4	<0.1
$\mathrm{M}(\varnothing)$	3.2	11.7	16.3	7.8	5.7	17.9	24.9	12.5
$\mathrm{M}(\mathcal{V})$	14.7	33.5	43.2	24.0	36.3	63.8	73.1	48.8
$\mathrm{M}(\mathcal{C}_{S_H})$	5.9	17.8	25.4	11.9	17.3	43.7	56.7	29.9
$\mathrm{M}(\mathcal{C}_{S_H}, \mathcal{V})$	15.0	34.7	44.7	24.7	41.6	70.6	78.6	54.2

Table 5.4 – Top-k ($k \in \{1, 5, 10\}$) (%) and MRR scores (%). $p_{\text{sup}} = 50\%$.

the final model $\mathrm{M}(\mathcal{C}_{S_H}, \mathcal{V})$ decreases after $p_{\text{sup}} = 50\%$. Since the highest relative improvement $\delta_{\mathcal{C}}$ (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.

5.4.2 Modeling contextual information

In this section, we compare the different context models; results are reported in Table 5.5. First, underlying hypotheses of our model are experimentally tested:

1. Modeling context and prior information with semantic representations (models $\mathrm{M}(\mathcal{C}_\star, \mathcal{V})$) 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.
2. Moreover, 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.* $\mathrm{M}(\mathcal{C}_{S_H}, \mathcal{V}) \overset{\text{MFR}}{<} \mathrm{M}(\mathcal{C}_{S_H} \oplus \mathcal{V})$).
3. Furthermore, we observe that our approach $\mathrm{M}(\mathcal{V})$ is more efficient to capture the imbalanced class distribution of the source domain, compared to DeViSE($\mathcal{V}$);

	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	DeViSE$(\mathcal{V})$	$P(.\|\mathcal{V})$	10.73	3.62
	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 5.5 – **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.

indeed, *True Prior* $\approx \text{M}(\mathcal{V})$, whereas *True Prior* $\overset{\text{MFR}}{<}$ DeViSE$(\mathcal{V})$ on $\mathcal{S}$. 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 5.4.1), using contextual information is always beneficial. Indeed, all models with context $\text{M}(\mathcal{C}_\star, \mathcal{V})$ improve over $\text{M}(\mathcal{V})$ — 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:

1. when the bounding boxes annotations are available: all of our models that use both $\mathcal{C}$ and $\mathcal{V}$ outperform the baseline $\text{M}(\mathcal{C}_I, \mathcal{V})$, 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,
2. 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 $S_H \cup T_H \overset{\text{MFR}}{<} S_H \cup T_L$,
3. when more context objects are considered (*e.g.* $S_L \cup T_L \overset{\text{MFR}}{<} S_L$),
4. when low-level information is used complementarily to high-level information (*e.g.* $S_L \cup S_H \cup T_L \overset{\text{MFR}}{<} S_L \cup T_L$).

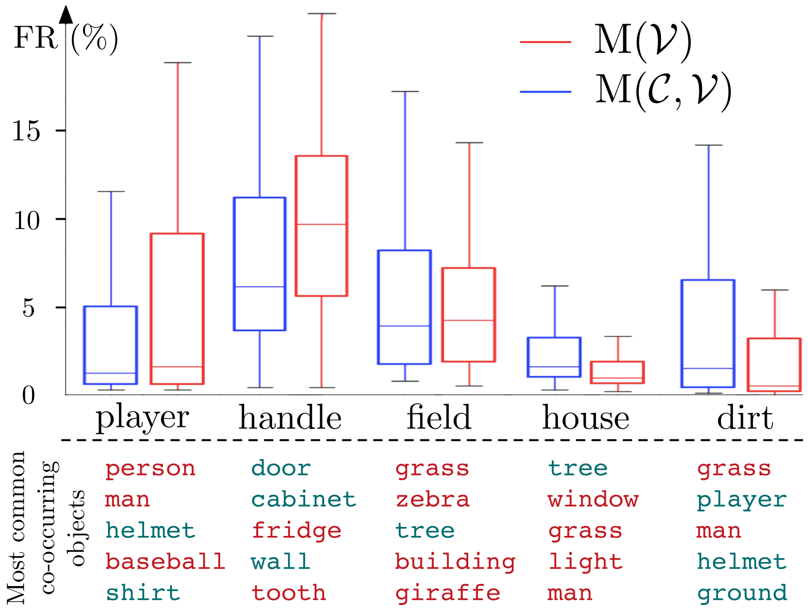


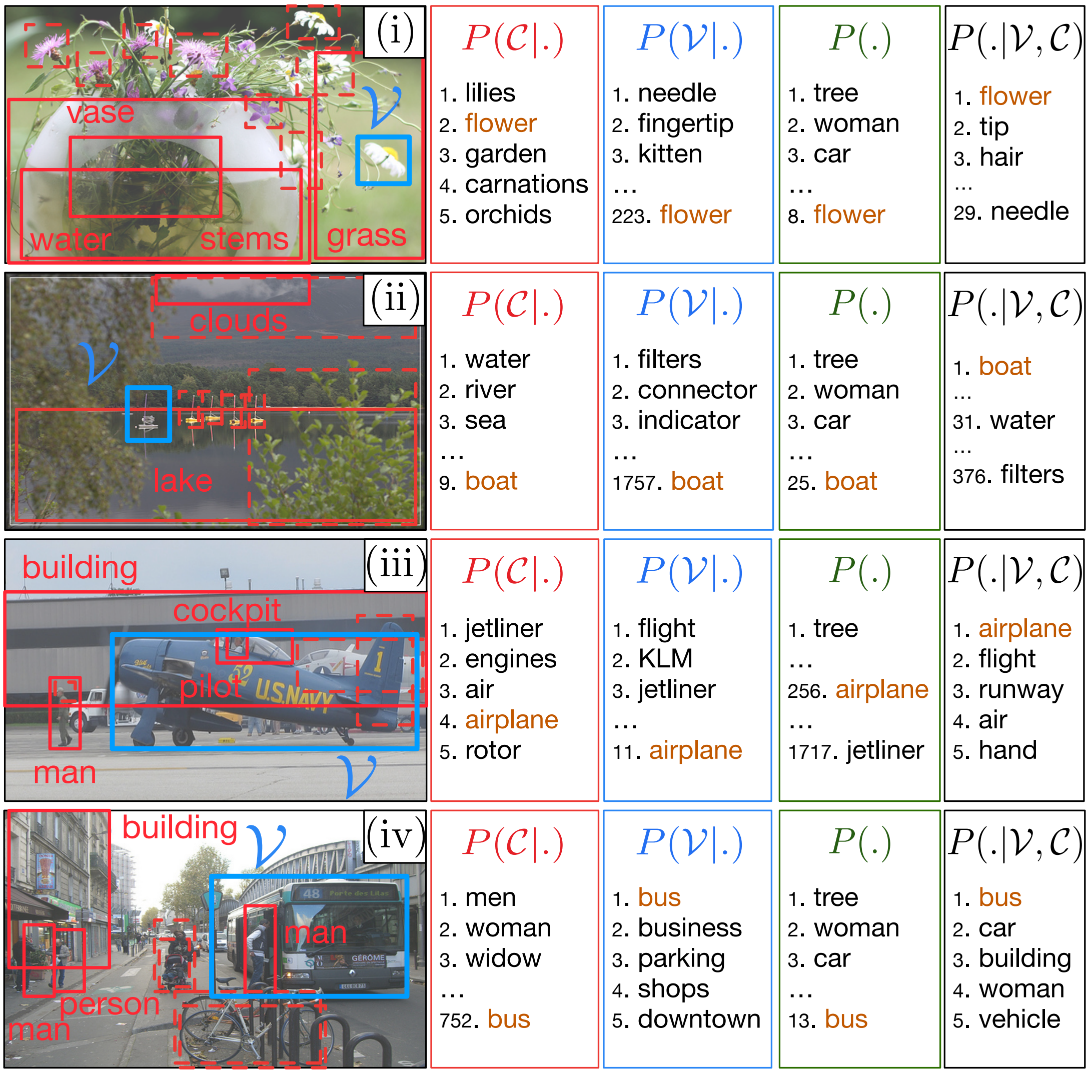
Figure 5.7 – **Qualitative analysis**. Boxplot representing the distribution of the correct ranks (First Relevant in %) for five randomly selected classes of the target domain, with the context model $S_L \cup S_H \cup T_L$. Below are listed, by order of frequency, the classes that co-occur the most with the object of interest (classes of $\mathcal{T}$ in green; $\mathcal{S}$ in red).

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% relative improvement in the target domain (32% in the 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.

5.4.3 Qualitative Experiments

To gain a deeper understanding of contextual information, we compare in Figure 5.7 the predictions of $\mathrm{M}(\mathcal{V})$ (the model without visual context) and the global model $\mathrm{M}(\mathcal{C}, \mathcal{V})$ which uses visual context. 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



	$P(\mathcal{C}\vert .)$	$P(\mathcal{V}\vert .)$	$P(.)$	$P(.\vert\mathcal{V},\mathcal{C})$
(i)	1. lilies 2. flower 3. garden 4. carnations 5. orchids	1. needle 2. fingertip 3. kitten … 223. flower	1. tree 2. woman 3. car … 8. flower	1. flower 2. tip 3. hair … 29. needle
(ii)	1. water 2. river 3. sea … 9. boat	1. filters 2. connector 3. indicator … 1757. boat	1. tree 2. woman 3. car … 25. boat	1. boat … 31. water … 376. filters
(iii)	1. jetliner 2. engines 3. air 4. airplane 5. rotor	1. flight 2. KLM 3. jetliner … 11. airplane	1. tree … 256. airplane … 1717. jetliner	1. airplane 2. flight 3. runway 4. air 5. hand
(iv)	1. men 2. woman 3. widow … 752. bus	1. bus 2. business 3. parking 4. shops 5. downtown	1. tree 2. woman 3. car … 13. bus	1. bus 2. car 3. building 4. woman 5. vehicle

Figure 5.8 – **Qualitative analysis: positive 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).

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*, *bat* . . .) that grounds *person* in a sport setting, and thus suggests that the *person* could be playing. Visual context is also particularly relevant when the object of interest has a generic shape. For example, *handle*, without context, is visually similar to many round objects; but it is the presence of objects like *door* or *fridge* in the context that helps determine the nature of the object of interest.

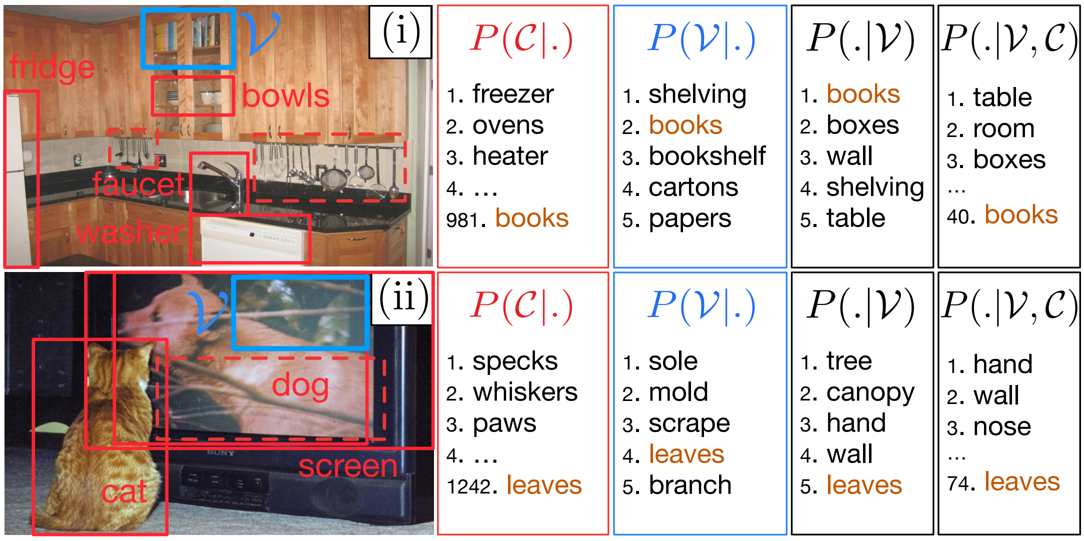


Figure 5.9 – **Qualitative analysis: negative examples** where the use of the context leads to degraded predictions, *i.e.* examples where model $\mathrm{M}(\mathcal{C}_{S_L \cup S_H \cup T_L}, \mathcal{V})$ is worse than the simpler model $\mathrm{M}(\mathcal{V})$ ($\mathcal{T}$ classes only).

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 5.8). 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, 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. Moreover, using contextual information can degrade predictions when an object occurs in an environment in which it is unexpected. For example, Figure 5.9 shows a picture of a kitchen where the object 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 (as it is unexpected). However, the model $\mathrm{M}(\mathcal{V})$ retrieves the correct label, given only the region of interest. Integrating contextual information in the final model $\mathrm{M}(\mathcal{C}_{S_L \cup S_H \cup T_L}, \mathcal{V})$ thus leads to worse performances over $\mathrm{M}(\mathcal{V})$.

5.5 Conclusion

5.5.1 Summary of the contributions

In this chapter, we showed that language can be used to measure the likelihood of an object within a particular visual context. Practically, we introduced a new approach for ZSL: *context-aware zero-shot learning*, along with a corresponding model which uses complementary contextual information to significantly improves predictions, both quantitatively and qualitatively, on the visual genome dataset, adapted for our setting.

Modeling the probability of visual context given a word lead us to reformulate the traditional ZSL task, and to separately model three components: the prior, the visual and the context components. By explicitly modeling the prior probability for both source and target classes and we found that the probability of appearance of an object in an image can be learned from linguistic data (*i.e.* with label representation). Moreover, our model is modular and interpretable: each component is learned separately and it is possible to measure the contributions brought by each module independently.

5.5.2 Perspectives

This work is a first step to show that language can be leveraged, along with visual contexts, to help visual understanding. Extensions and future work could include the following perspectives.

Towards a real-world scenario In this chapter, a strong assumption is made: bounding boxes are provided for all objects (from both the source and target domains) on both the training and testing sets. However, in the real world, it seems unlikely to have access to bounding boxes for objects of the target domain as annotating bounding boxes is more time consuming than collecting labels.

To make our work fully applicable to a real-world scenario, removing the need for pre-detected object boxes is then a necessity. In practice, this could be made possible with the use of algorithms that localize objects, without the need of supervision. Examples of such algorithms include Selective Search (Uijlings et al. 2013), EdgeBox (Zitnick et al. 2014) or the Region Proposer Network (S. Ren et al. 2017).

Spatial features of objects Images intrinsically contain spatial information on how are objects located in a scene — we exploited this intuition to obtained improved embeddings representing word semantics in Section 3.3.1.3. To some extent, this spatial knowledge also appears in language data (Kordjamshidi et al.

2011). For example, one could read in a text that a keyboard lies next to a monitor and a mouse. Such spatial common-sense could be acquired from language, such as with within the spatial role labeling task (Kordjamshidi et al. 2010), to further facilitate visual recognition of an object given the spatial arrangement of the scene, as well as its context as explored in this chapter.

Such an approach relies on the ability to extract spatial common-sense from language data which is an open challenge as language is lacks common-sense as as explained in Chapter 3. Existing work, such as (Collell et al. 2018a), show that it is possible to learn spatial arrangements for words and relations, even in if they have never been encountered in images.

Using grounded word embeddings Designing grounded word embeddings that include more visual contextual information would greatly benefit our model, and especially the context component. However, it would not be possible to learn such word embeddings with visual data for all concepts that appear in images, as done in Chapter 3, as it requires supervision for all words which is impossible in a zero-shot setting. An open question thus remains: given textual data only, how to learn semantic representation from which visual co-occurrence statistics can be recovered?

Chapter 6

CONCLUSION

Contents

6.1 Summary of Contributions

In this thesis, we tackled challenging problems of multimodal machine learning dealing with language and visual data. We undertook two complementary approaches, and showed that leveraging one modality (text or image) can benefit the other one (image or text). Our contributions can be organized in two axes detailed below.

Grounding language in the visual world
Representing the semantics of words and sentences is a long-standing problem for the Natural Language Processing (NLP) community, and most methods build word semantic representations given their textual context in large corpora (such as with the `skip-gram` and `SkipThought` algorithms, for words and sentences respectively). However, it is widely known that linguistic statistics are different from real-world statistics, for example in terms of word occurrence or co-occurrence frequencies. This has the effect of producing representations that are biased, that lack perceptual information and common-sense knowledge. A recent line of research, which fits the scope of our contributions in Chapter 3 and Chapter 4, attempt to *ground* language and more precisely to improve textual representations by additionally integrating visual features.

In the case of words (Chapter 3), we hypothesized that semantic information about an object is contained both in (1) the visual surrounding of that object in an image, and (2) the spatial organization of the other objects in the image. We thus proposed to incorporate visual context of objects, and spatial information, into semantic representations, and thus designed a model based on the `skip-gram` algorithm, where word representations are learned and trained to predict their textual and visual contexts. We found that visual surroundings of objects, and their relative spatial organization, are very informative to build word representations. Indeed, leveraging this visual information improves word representations over traditional approaches which only use the visual appearance of objects themselves (+2. on average on word evaluation benchmarks). This work has been published in AAAI 2018. Our code for evaluating word representation has been open-sourced [1] as well as obtained grounded word representations [2].

In the case of sentences (Chapter 4), other challenges were involved as sentences can be visually ambiguous, carry non-visual information, or have a wide variety of paraphrases and related sentences describing a same scene. To deal with these perspectives, we proposed to transfer visual information to textual representations by defining an intermediate representation space: the grounded space. This space allows us to define two complementary objectives that we can then optimize without over-constraining the textual space. These objectives ensure that (1) sentences associated with the same visual content are close in the grounded space and that (2) similarities between related elements are preserved across modalities. We showed both quantitatively and qualitatively that grounding brings useful complementary information, to both concrete and more abstract sentences. Besides, the approach is the first to report consistent positive results against purely textual baselines on a variety of natural language tasks (+1.3 on average on SentEval). The work is under review at EMNLP 2019.

Leveraging language for visual understanding
Images can be used to ground word and sentence semantics, and conversely, leveraging natural language can also benefit visual recognition. Indeed, natural language can be used to help computer vision systems, either to evaluate visual reasoning capacities (Section 2.3.1), or to augment capacities of computer vision recognition systems, especially when visual supervision is limited (Section 2.3.2). In the latter case, visual systems can benefit from high-level semantic word representations, which are learned from purely textual resources. Typically, these representation encode some information about the visual appearance of objects and this information is crucial to recognize objects for which no visual supervision is given during training. This scenario corresponds to the zero-shot recognition task. In Chapter 5, we questioned whether semantic representations encode other kinds of visual information, beyond the visual appearance knowledge: is information about (1) possible visual context of objects, and about (2) visual occurrence likelihood, contained within word representations? To answer these questions, we extended the zero-shot recognition task to additionally leverage visual context of unseen objects, this constitutes the new task: *context-aware zero-shot recognition*. To explicitly model and measure contextual information and investigate its complementarity to information contained within the area of interest, we cast the

[1] github.com/EloiZ/embedding_evaluation

[2] data.lip6.fr/multimodal_embeddings/

zero-shot learning problem into a bayesian formalism. We present a model for zero-shot recognition that leverages (1) the region of interest, (2) the semantic representations of objects, and (3) the visual context of an object. Moreover, we proposed efficient sampling strategies to learn the proposed model in this adapted formulation, and to take into account imbalanced class distributions. To conduct experiments, we designed a new dataset, based on the *visual genome* dataset, which has been open-sourced[3] to facilitate future works on context-aware zero-shot learning. We found that information about visual occurrence likelihood is contained within object representations, computed from textual data. Besides, we found that information about possible visual environments of objects is contained within representations, and that using this information leads to a 22% relative improvement over other state-of-the-art zero-shot recognition systems (on a ranking metric). Finally, we conducted extensive quantitative and qualitative evaluations to gain a deeper understanding on how visual context impacts zero-shot recognition, and found that, for example, context can help refining predictions of specific classes and disambiguating generic shapes. Overall, the use of the visual context gives a relative improvement of 22% (Mean First Relevant (MFR) metric) over the baseline model which ignores visual surroundings. This work and results have been published at ICML 2019.

6.2 Open questions and perspectives

Starting from our contributions, several questions remain and we foresee some perspectives to seamlessly handle both visual and textual worlds. We split these perspectives into three categories: (1) follow-up extensions of our approaches, (2) other perspectives that can immediately be tackled, regarding the background given in Chapter 2, and (3) more ambitious and long-term perspectives which require more careful thinking but could have bigger impacts.

6.2.1 Extensions and perspectives of our approaches

Incorporating visual semantics to linguistic representations, detailed in Section 3.6.2 and in Section 4.5.2. To further incorporate common-sense semantics and real-world knowledge to word representations, one can possibly leverage other information sources than images. This includes the use of dictionary definitions (Tissier et al. 2017), Knowledge Base (KB) (Weston et al. 2013; Shalaby et al. 2018), audio signals (Kiela et al. 2017) or olfactory knowledge (Kiela et al. 2015a).

Moreover, a possible extension is to learn multimodal representations for *relations* of the form (*subject*, *predicate*, *object*). In particular, the semantics contained

[3] data.lip6.fr/context_aware_zsl/

in the visual context of the relation could further improve the quality of the relation embeddings, for example by incorporating information about the typical environments in which the relation occurs.

Finally, we observed that there is no straightforward way to quantitatively and qualitatively measure visual information contained within textual embeddings (for both words and sentences). Finding meaningful ways to explore the in-depth difference between various representation spaces (*e.g.* non-grounded vs. grounded ones) remains an open question.

Textual representations for the zero-shot recognition task, detailed in Section 5.5.2. In Chapter 5, we suppose that bounding boxes are known for all objects (from both the source and target domains) on both the training and testing sets. However, in the real world, it seems unlikely to have access to bounding boxes for objects of the target domain as annotating bounding boxes is more time consuming than collecting labels at the image labels. To make our work fully applicable to a real-world scenario, removing the need for pre-detected object boxes is then a necessity. In practice, this could be made possible with the use of algorithms that localize objects, without the need of supervision. Examples of such algorithms include Selective Search (Uijlings et al. 2013), EdgeBox (Zitnick et al. 2014) or the Region Proposer Network (S. Ren et al. 2017).

To some extent, spatial knowledge appears in language data (Kordjamshidi et al. 2011), as for example language can state that a 'keyboard' can be found *next to* a 'computer mouse' and *below* a 'monitor'. Provided that this spatial knowledge is correctly encoded in semantic textual representations, one could then leverage this spatial knowledge to further boost the zero-shot recognition task. The very same reasoning could apply with temporal knowledge and the zero-shot action recognition task in videos.

6.2.2 Research perspectives

Learn to make semantically more plausible errors. Using linguistic data to augment capacities of visual recognition systems has so far been limited to cases where supervision is scarce, such as with the Visual Relationship Detection (VRD) and zero-shot recognition tasks. When data is abundant, another research direction would be to learn visual systems that make semantically more plausible errors, thanks to linguistic knowledge. This idea has been explored in some specific domains such as food recognition systems (H. Wu et al. 2016) and fine-grain image classification (Zhang et al. 2016). This can be achieved by learning deep hierarchies of concepts from texts and by penalizing more or less models, given distinctions between semantically incorrect mistakes and less serious ones.

Language/vision multi-task learning. We hypothesize that jointly solving multiple tasks could yield increased performances over the approach which independently solves each task. Typically, when inputs are composed of multimodal channels (*e.g.* a movie and its subtitles, an image and its caption), ambiguities in a modality can be alleviated by leveraging the complementary modality. For example, Ramanathan et al. 2014 consider the problems of *person naming* in videos (to give a name to a track of a person in a video) and *coreference resolution* in text. They develop a joint model for these tasks that infer a latent alignment between tracks and mentions and show significant improvement over the independent baselines.

Besides, Christie et al. 2017 consider the problems of *semantic segmentation* in images and *prepositional phrase attachment resolution* in sentences paired with those images. Given one of the task, they assume to have a model that outputs scores for potential solutions and a list a plausible (high-scoring) and non-redundant (diverse) hypotheses of resolution. At inference, a factor graph assesses the consistency of the simultaneous choice on the two tasks and yields improved results over the independent baselines.

Exploit *comic strip* data. Comic strips are complex data involving text, drawings, panels, speech bubbles, and onomatopoeia. A high-level analysis is necessary to fully understand events, emotions, storytelling, actions, drawings, and the relations between characters.

On the one hand, comics contain high-level visual information in the form of stylized pictures that are visually different to photographs, while sharing similar content. Then, from a computer vision perspective, comics could be used to design transfer learning and domain adaptation approaches, for applications such as unsupervised object detection and segmentation.

On the other hand, surprisingly few approaches have been proposed to analyze the storyline based on the text. From a NLP perspective, we assume that the order of the panels, the way the story is cut, the choice of the content is directly related to the way humans reason and talk. Exploiting this information could help to learn common-sense knowledge and could thus be useful for downstream NLP applications. Working with comic strip data is ambitious and can lead to answers to fundamental questions at the interplay of language and computer vision.

6.2.3 Longer-term research directions

Model and quantify the human reporting bias. Is it possible to model the bias between the content of images and natural language? In this thesis, we exploited the complementarity of language and vision, to benefit tasks in each modalities. However, we still lack concrete understanding about the nature of the human reporting bias, beyond the original observations by Gordon et al. 2013 detailed

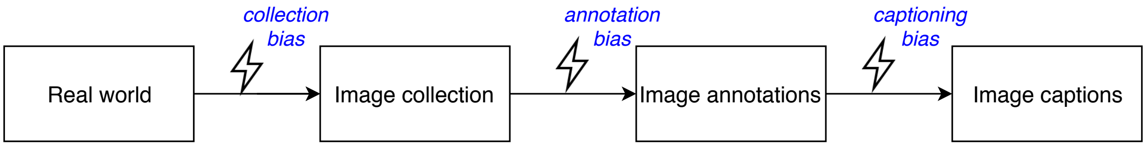


Figure 6.1 – **Linguistic and Visual bias**. The *collection bias* between the real world and image collections is introduced by the way images are selected (for example, important objects are usually near the center of the image). As it is not possible to exhaustively annotate every objects within an image, the *annotation bias* emerges (*e.g.* salient and "important" objects are more likely to be labelled). Given an image, and a list of annotated objects, the *captioning bias* emerges when human generate a caption (for example, only a few objects will be mentioned in the caption)

in Section 2.2.1. We argue that this bias has multiple origins, and we attempt to illustrate some in Figure 6.1 (captions generated from images). The global understanding of these biases remains an open question, and we believe that answers to this question will have crucial impacts in the scope of multimodal machine learning with text and image, and more generally in the NLP field. The path towards this goal is *a priori* not exclusively contained within the fields of computer science and computational linguistics, and other domains should be involved such as sociology and psycholinguistics.

Leverage visual knowledge for real-world NLP tasks. So far using visual knowledge for NLP applications has shown benefits on intrinsic evaluation of word/sentence representations (as reviewed in Section 2.2.2 and shown in Chapter 3 and Chapter 4), and on common-sense tasks. While common-sense tasks are good playgrounds to evaluate grounded models, these tasks are toyish and transfer capacities of grounded models to real-world/downstream tasks remains unclear. To the best of our knowledge, state-of-the-art models for automatic translation, open-domain question-answering and language modelling are based on purely textual data. A research direction is then to integrate visual knowledge to these models.

BIBLIOGRAPHY

Abadi, Martin, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. (2016). "Tensorflow: Large-scale machine learning on heterogeneous distributed systems". In: *arXiv preprint arXiv:1603.04467* (cit. on p. 57).

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. URL: http://www.aaai.org/Library/Symposia/Spring/2005/ss05-03-014.php (cit. on p. 27).

Akata, Zeynep, Florent Perronnin, Zaıd Harchaoui, and Cordelia Schmid (2016). "Label-Embedding for Image Classification". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 38.7, pp. 1425–1438. URL: https://doi.org/10.1109/TPAMI.2015.2487986 (cit. on pp. 43, 81).

Akata, Zeynep, Scott E. Reed, Daniel Walter, Honglak Lee, and Bernt Schiele (2015). "Evaluation of output embeddings for fine-grained image classification". In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 2927–2936. URL: https://doi.org/10.1109/CVPR.2015.7298911 (cit. on p. 43).

Akbik, Alan and Thilo Michael (2014). "The Weltmodell: A Data-Driven Commonsense Knowledge Base". In: *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, Reykjavik, Iceland, May 26-31, 2014*. Pp. 3272–3276. URL: http://www.lrec-conf.org/proceedings/lrec2014/summaries/409.html (cit. on p. 28).

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. URL: http://jmlr.org/proceedings/papers/v28/andrew13.html (cit. on p. 24).

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. URL: http://arxiv.org/abs/1904.00688 (cit. on p. 11).

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. URL: https://doi.org/10.1016/j.csl.2019.01.005 (cit. on p. 17).

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*. URL: https://openreview.net/forum?id=Sy2ogebAW (cit. on p. 11).

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. URL: https://doi.org/10.1109/ICCV.2015.483 (cit. on p. 43).

Bach, Francis R. and Michael I. Jordan (2002). "Kernel Independent Component Analysis". In: *Journal of Machine Learning Research* 3, pp. 1–48. URL: http://jmlr.org/papers/v3/bach02a.html (cit. on p. 24).

Bagherinezhad, Hessam, Hannaneh Hajishirzi, Yejin Choi, and Ali Farhadi (2016). "Are Elephants Bigger than Butterflies? Reasoning about Sizes of Objects". In: *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA.* Pp. 3449–3456. URL: http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/12510 (cit. on pp. 27, 28).

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2015). "Neural Machine Translation by Jointly Learning to Align and Translate". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. URL: http://arxiv.org/abs/1409.0473 (cit. on pp. 1, 11, 14, 16, 64).

Bakarov, Amir (2018). "A Survey of Word Embeddings Evaluation Methods". In: *CoRR* abs/1801.09536. arXiv: 1801.09536. URL: http://arxiv.org/abs/1801.09536 (cit. on pp. 32, 33).

Bansal, Ankan, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran (2018). "Zero-Shot Object Detection". In: *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part I*, pp. 397–414. URL: https://doi.org/10.1007/978-3-030-01246-5%5C_24 (cit. on pp. 43, 85).

Baroni, Marco (2016). "Grounding Distributional Semantics in the Visual World". In: *Language and Linguistics Compass* 10.1, pp. 3–13. URL: https://doi.org/10.1111/lnc3.12170 (cit. on pp. 14, 27).

Baroni, Marco, Georgiana Dinu, and Germán Kruszewski (2014). "Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors". 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. 238–247. URL: http://aclweb.org/anthology/P/P14/P14-1023.pdf (cit. on p. 14).

Barsalou, Lawrence W. (Jan. 2008). “Grounded Cognition”. In: *Annual Review of Psychology* 59.1, pp. 617–645. URL: http://www.annualreviews.org/doi/10.1146/annurev.psych.59.103006.093639 (cit. on pp. 27, 30, 46).

Bell, Sean, C. Lawrence Zitnick, Kavita Bala, and Ross B. Girshick (2016). “Inside-Outside Net: Detecting Objects in Context with Skip Pooling and Recurrent Neural Networks”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 2874–2883. URL: https://doi.org/10.1109/CVPR.2016.314 (cit. on p. 81).

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. URL: https://doi.org/10.1109/ICCV.2017.285 (cit. on pp. 22, 40).

Ben-younes, Hedi, Rémi Cadène, Nicolas Thome, and Matthieu Cord (2019). “BLOCK: Bilinear Superdiagonal Fusion for Visual Question Answering and Visual Relationship Detection”. In: *CoRR* abs/1902.00038. arXiv: 1902.00038. URL: http://arxiv.org/abs/1902.00038 (cit. on p. 22).

Bengio, Samy, Jeffrey Dean, Dumitru Erhan, Eugene Ie, Quoc V. Le, Andrew Rabinovich, Jonathon Shlens, and Yoram Singer (2013). “Using Web Co-occurrence Statistics for Improving Image Categorization”. In: *CoRR* abs/1312.5697. arXiv: 1312.5697. URL: http://arxiv.org/abs/1312.5697 (cit. on p. 93).

Bengio, Yoshua (2008). “Neural net language models”. In: *Scholarpedia* 3.1, p. 3881. URL: https://doi.org/10.4249/scholarpedia.3881 (cit. on p. 14).

Bengio, Yoshua, Aaron C. Courville, and Pascal Vincent (2013). “Representation Learning: A Review and New Perspectives”. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 35.8, pp. 1798–1828. URL: https://doi.org/10.1109/TPAMI.2013.50 (cit. on pp. 1, 10).

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Janvin (2003). “A Neural Probabilistic Language Model”. In: *Journal of Machine Learning Research* 3, pp. 1137–1155. URL: http://jmlr.org/papers/v3/bengio03a.html (cit. on pp. 1, 13, 14).

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. URL: https://doi.org/10.1109/72.279181 (cit. on p. 13).

Bergsma, Shane and Benjamin Van Durme (2011a). “Learning Bilingual Lexicons Using the Visual Similarity of Labeled Web Images”. In: *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011*, pp. 1764–1769. URL: https://doi.org/10.5591/978-1-57735-516-8/IJCAI11-296 (cit. on p. 37).

Bergsma, Shane and Randy Goebel (2011b). “Using Visual Information to Predict Lexical Preference”. In: *Recent Advances in Natural Language Processing, RANLP*

2011, 12-14 September, 2011, Hissar, Bulgaria, pp. 399–405. URL: http://www.aclweb.org/anthology/R11-1055 (cit. on p. 36).

Berzak, Yevgeni, Andrei Barbu, Daniel Harari, Boris Katz, and Shimon Ullman (2015). "Do You See What I Mean? Visual Resolution of Linguistic Ambiguities". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pp. 1477–1487. URL: http://aclweb.org/anthology/D/D15/D15-1172.pdf (cit. on p. 36).

Blei, David M., Andrew Y. Ng, and Michael I. Jordan (2001). "Latent Dirichlet Allocation". In: *Advances in Neural Information Processing Systems 14 [Neural Information Processing Systems: Natural and Synthetic, NIPS 2001, December 3-8, 2001, Vancouver, British Columbia, Canada]*, pp. 601–608. URL: http://papers.nips.cc/paper/2070-latent-dirichlet-allocation (cit. on p. 18).

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. URL: http://papers.nips.cc/paper/5071-translating-embeddings-for-modeling-multi-relational-data (cit. on p. 61).

Bordes, Patrick, Éloi Zablocki, Laure Soulier, Benjamin Piwowarski, and Patrick Gallinari (2019). "Incorporating Visual Semantics into Sentence Representations within a Grounded Space". In: *EMNLP 2019* (cit. on pp. 6, 64).

Borenstein, Eran and Jitendra Malik (2006). "Shape Guided Object Segmentation". In: *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), 17-22 June 2006, New York, NY, USA*, pp. 969–976. URL: https://doi.org/10.1109/CVPR.2006.276 (cit. on p. 2).

Boutonnet, Bastien and Gary Lupyan (2015). "Words jump-start vision: A label advantage in object recognition". In: *Journal of Neuroscience* 35.25, pp. 9329–9335 (cit. on p. 27).

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. URL: http://aclweb.org/anthology/D/D15/D15-1075.pdf (cit. on p. 35).

Bruni, Elia, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran (2012). "Distributional Semantics in Technicolor". In: *The 50th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, July 8-14, 2012, Jeju Island, Korea - Volume 1: Long Papers*, pp. 136–145. URL: http://www.aclweb.org/anthology/P12-1015 (cit. on pp. 27, 30, 47).

Bruni, Elia, Nam-Khanh Tran, and Marco Baroni (2014). "Multimodal Distributional Semantics". In: *J. Artif. Intell. Res.* 49, pp. 1–47. URL: https://doi.org/10.1613/jair.4135 (cit. on pp. 30, 32, 46–48).

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. 72).

Bucher, Maxime, Stéphane Herbin, and Frédéric Jurie (2016). "Improving Semantic Embedding Consistency by Metric Learning for Zero-Shot Classiffication". In: *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V*, pp. 730–746. URL: https://doi.org/10.1007/978-3-319-46454-1%5C_44 (cit. on p. 43).

Buscaldi, Davide and Paolo Rosso (2006). "A Naıve Bag-of-Words Approach to Wikipedia QA". In: *Working Notes for CLEF 2006 Workshop co-located with the 10th European Conference on Digital Libraries (ECDL 2006), Alicante, Spain, September 20-22, 2006.* URL: http://ceur-ws.org/Vol-1172/CLEF2006wn-QACLEF-BuscaldiEt2006b.pdf (cit. on p. 12).

Cadène, Rémi, Hedi Ben-younes, Matthieu Cord, and Nicolas Thome (2019). "MUREL: Multimodal Relational Reasoning for Visual Question Answering". In: *CoRR* abs/1902.09487. arXiv: 1902.09487. URL: http://arxiv.org/abs/1902.09487 (cit. on p. 40).

Caglayan, Ozan, Walid Aransa, Adrien Bardet, Mercedes Garcıa-Martınez, Fethi Bougares, Loıc Barrault, Marc Masana, Luis Herranz, and Joost van de Weijer (2017). "LIUM-CVC Submissions for WMT17 Multimodal Translation Task". In: *Proceedings of the Second Conference on Machine Translation, WMT 2017, Copenhagen, Denmark, September 7-8, 2017*, pp. 432–439. URL: https://aclanthology.info/papers/W17-4746/w17-4746 (cit. on p. 37).

Caglayan, Ozan, Walid Aransa, Yaxing Wang, Marc Masana, Mercedes Garcıa-Martınez, Fethi Bougares, Loıc Barrault, and Joost van de Weijer (2016). "Does Multimodality Help Human and Machine for Translation and Image Captioning?" In: *Proceedings of the First Conference on Machine Translation, WMT 2016, colocated with ACL 2016, August 11-12, Berlin, Germany*, pp. 627–633. URL: http://aclweb.org/anthology/W/W16/W16-2358.pdf (cit. on p. 37).

Carlson, Andrew, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr., and Tom M. Mitchell (2010). "Toward an Architecture for Never-Ending Language Learning". In: *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, USA, July 11-15, 2010.* URL: http://www.aaai.org/ocs/index.php/AAAI/AAAI10/paper/view/1879 (cit. on p. 28).

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. URL: https://doi.org/10.1145/3209978.3210036 (cit. on pp. 25, 67).

Castrejon, Lluis, Yusuf Aytar, Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba (2016). "Learning Aligned Cross-Modal Representations from Weakly Aligned Data". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 2940–2949. URL: https://doi.org/10.1109/CVPR.2016.321 (cit. on p. 73).

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. URL: https://doi.org/10.18653/v1/S17-2001 (cit. on p. 35).

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. URL: https://aclanthology.info/papers/D18-2029/d18-2029 (cit. on p. 17).

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. URL: https://doi.org/10.18653/v1/P17-1171 (cit. on p. 11).

Chen, Minmin (2017). "Efficient Vector Representation for Documents through Corruption". In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. URL: https://openreview.net/forum?id=B1Igu2ogg (cit. on p. 18).

Cho, Kyunghyun, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio (2014). "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation". 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. 1724–1734. URL: http://aclweb.org/anthology/D/D14/D14-1179.pdf (cit. on pp. 13, 17).

Christie, Gordon, Ankit Laddha, Aishwarya Agrawal, Stanislaw Antol, Yash Goyal, Kevin Kochersberger, and Dhruv Batra (2017). "Resolving vision and language ambiguities together: Joint segmentation & prepositional attachment resolution in captioned scenes". In: *Computer Vision and Image Understanding* 163, pp. 101–112. URL: https://doi.org/10.1016/j.cviu.2017.09.001 (cit. on pp. 36, 107).

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. URL: http://aclweb.org/anthology/P/P15/P15-2019.pdf (cit. on p. 33).

Collell, Guillem, Luc Van Gool, and Marie-Francine Moens (2018a). "Acquiring Common Sense Spatial Knowledge Through Implicit Spatial Templates". 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. 6765–6772. URL: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16232 (cit. on p. 102).

Collell, Guillem and Marie-Francine Moens (2016). "Is an Image Worth More than a Thousand Words? On the Fine-Grain Semantic Differences between Visual and Linguistic Representations". In: *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*, pp. 2807–2817. URL: http://aclweb.org/anthology/C/C16/C16-1264.pdf (cit. on pp. 32, 48, 58).

Collell, Guillem and Marie-Francine Moens (2018b). "Do Neural Network Cross-Modal Mappings Really Bridge Modalities?" In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 2: Short Papers*, pp. 462–468. URL: https://aclanthology.info/papers/P18-2074/p18-2074 (cit. on pp. 64, 71, 72).

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. URL: http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14811 (cit. on pp. 30, 31, 57, 64, 70, 75, 76).

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. 35, 75).

Conneau, Alexis, Douwe Kiela, Holger Schwenk, Loıc Barrault, and Antoine Bordes (2017). "Supervised Learning of Universal Sentence Representations from Natural Language Inference Data". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pp. 670–680. URL: https://aclanthology.info/papers/D17-1070/d17-1070 (cit. on pp. 2, 17).

Dai, Andrew M., Christopher Olah, and Quoc V. Le (2015). "Document Embedding with Paragraph Vectors". In: *CoRR* abs/1507.07998. arXiv: 1507.07998. URL: http://arxiv.org/abs/1507.07998 (cit. on p. 18).

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. URL: https://doi.org/10.1109/CVPR.2005.177 (cit. on pp. 2, 18).

Das, Abhishek, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra (2017). "Visual Dialog". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 1080–1089. URL: https://doi.org/10.1109/CVPR.2017.121 (cit. on p. 40).

Deena, Salil, Madina Hasan, Mortaza Doulaty, Oscar Saz, and Thomas Hain (2019). "Recurrent Neural Network Language Model Adaptation for Multi-Genre Broadcast Speech Recognition and Alignment". In: *IEEE/ACM Trans. Audio, Speech & Language Processing* 27.3, pp. 572–582. URL: https://doi.org/10.1109/TASLP.2018.2888814 (cit. on p. 11).

Deerwester, Scott C., Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman (1990). "Indexing by Latent Semantic Analysis". In: *JASIS* 41.6, pp. 391–407. URL: https://doi.org/10.1002/(SICI)1097-4571(199009)41:6%5C%3C391::AID-ASI1%5C%3E3.0.CO;2-9 (cit. on p. 1).

Demirel, Berkan, Ramazan Gokberk Cinbis, and Nazli Ikizler-Cinbis (2018). "Zero-Shot Object Detection by Hybrid Region Embedding". In: *British Machine Vision Conference 2018, BMVC 2018, Northumbria University, Newcastle, UK, September 3-6, 2018*, p. 56. URL: http://bmvc2018.org/contents/papers/0136.pdf (cit. on p. 43).

Deng, Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li (2009). "ImageNet: A large-scale hierarchical image database". In: *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pp. 248–255. URL: https://doi.org/10.1109/CVPRW.2009.5206848 (cit. on pp. 20, 80, 94).

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2018). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *CoRR* abs/1810.04805. arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805 (cit. on pp. 2, 14, 16).

Divvala, Santosh Kumar, Ali Farhadi, and Carlos Guestrin (2014). "Learning Everything about Anything: Webly-Supervised Visual Concept Learning". In: *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pp. 3270–3277. URL: https://doi.org/10.1109/CVPR.2014.412 (cit. on pp. 27, 36).

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*. URL: http://www.aclweb.org/anthology/C04-1051 (cit. on p. 35).

Downey, Doug, Oren Etzioni, and Stephen Soderland (2005). “A Probabilistic Model of Redundancy in Information Extraction”. In: *IJCAI-05, Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence, Edinburgh, Scotland, UK, July 30 - August 5, 2005*, pp. 1034–1041. URL: http://ijcai.org/Proceedings/05/Papers/1390.pdf (cit. on p. 28).

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. URL: https://tel.archives-ouvertes.fr/tel-01635374 (cit. on p. 20).

Elman, Jeffrey L. (1990). “Finding Structure in Time”. In: *Cognitive Science* 14.2, pp. 179–211. URL: https://doi.org/10.1207/s15516709cog1402%5C_1 (cit. on p. 13).

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. URL: http://openaccess.thecvf.com/content%5C_cvpr%5C_2018/html/Engilberge%5C_Finding%5C_Beans%5C_in%5C_CVPR%5C_2018%5C_paper.html (cit. on p. 39).

Etzioni, Oren, Michele Banko, Stephen Soderland, and Daniel S. Weld (2008). “Open information extraction from the web”. In: *Commun. ACM* 51.12, pp. 68–74. URL: https://doi.org/10.1145/1409360.1409378 (cit. on p. 28).

Farhadi, Ali, Ian Endres, Derek Hoiem, and David A. Forsyth (2009). “Describing objects by their attributes”. In: *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pp. 1778–1785. URL: https://doi.org/10.1109/CVPRW.2009.5206772 (cit. on pp. 43, 80, 90).

Faruqui, Manaal, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer (2016). “Problems With Evaluation of Word Embeddings Using Word Similarity Tasks”. In: *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP, RepEval@ACL 2016*. URL: https://doi.org/10.18653/v1/W16-2506 (cit. on pp. 33, 48).

Feng, Yang, Lin Ma, Wei Liu, and Jiebo Luo (2018). “Unsupervised Image Captioning”. In: *CoRR* abs/1811.10787. arXiv: 1811.10787. URL: http://arxiv.org/abs/1811.10787 (cit. on p. 39).

Ferrante, Marco, Nicola Ferro, and Silvia Pontarollo (2017). “Are IR Evaluation Measures on an Interval Scale?” In: *Proceedings of the ACM SIGIR International*

Conference on Theory of Information Retrieval, ICTIR 2017, Amsterdam, The Netherlands, October 1-4, 2017, pp. 67–74. URL: https://doi.org/10.1145/3121050.3121058 (cit. on p. 91).

Ferreira, Fernanda and Michael K Tanenhaus (2007). "Introduction to the special issue on language–vision interactions". In: *Journal of Memory and Language* 57.4, pp. 455–459 (cit. on p. 27).

Finkelstein, Lev, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin (2002). "Placing search in context: the concept revisited". In: *ACM Trans. Inf. Syst.* 20.1, pp. 116–131. URL: https://doi.org/10.1145/503104.503110 (cit. on p. 32).

Frinken, Volkmar, Francisco Zamora-Martınez, Salvador España Boquera, Marıa José Castro Bleda, Andreas Fischer, and Horst Bunke (2012). "Long-short term memory neural networks language modeling for handwriting recognition". In: *Proceedings of the 21st International Conference on Pattern Recognition, ICPR 2012, Tsukuba, Japan, November 11-15, 2012*, pp. 701–704. URL: http://ieeexplore.ieee.org/document/6460231/ (cit. on p. 11).

Frome, Andrea, Gregory S. Corrado, Jonathon Shlens, Samy Bengio, Jeffrey Dean, Marc'Aurelio Ranzato, and Tomas Mikolov (2013). "DeViSE: A Deep Visual-Semantic Embedding Model". 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. 2121–2129. URL: http://papers.nips.cc/paper/5204-devise-a-deep-visual-semantic-embedding-model (cit. on pp. 5, 43, 84, 85, 87, 89, 94).

Fu, Yanwei, Timothy M. Hospedales, Tao Xiang, and Shaogang Gong (2014). "Learning Multimodal Latent Attributes". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 36.2, pp. 303–316. URL: https://doi.org/10.1109/TPAMI.2013.128 (cit. on p. 43).

Fu, Yanwei, Yongxin Yang, Timothy M. Hospedales, Tao Xiang, and Shaogang Gong (2015). "Transductive Multi-label Zero-shot Learning". In: *CoRR* abs/1503.07790. arXiv: 1503.07790. URL: http://arxiv.org/abs/1503.07790 (cit. on p. 80).

Fu, Zhen-Yong, Tao A. Xiang, Elyor Kodirov, and Shaogang Gong (2015). "Zero-shot object recognition by semantic manifold distance". In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 2635–2644. URL: https://doi.org/10.1109/CVPR.2015.7298879 (cit. on p. 43).

Fuhr, Norbert (2017). "Some Common Mistakes In IR Evaluation, And How They Can Be Avoided". In: *SIGIR Forum* 51.3, pp. 32–41. URL: https://doi.org/10.1145/3190580.3190586 (cit. on p. 91).

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. URL: https://doi.org/10.1016/0031-3203(82)90024-3 (cit. on pp. 2, 19).

Glenberg, Arthur M and Michael P Kaschak (2002). "Grounding language in action". In: *Psychonomic bulletin & review* (cit. on pp. 27, 30, 46).

Glorot, Xavier, Antoine Bordes, and Yoshua Bengio (2011). "Deep Sparse Rectifier Neural Networks". In: *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011, Fort Lauderdale, USA, April 11-13, 2011*, pp. 315–323. URL: http://proceedings.mlr.press/v15/glorot11a/glorot11a.pdf (cit. on p. 20).

Goldberg, Yoav and Omer Levy (2014). "word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method". In: *CoRR* abs/1402.3722. arXiv: 1402.3722. URL: http://arxiv.org/abs/1402.3722 (cit. on p. 16).

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. URL: https://doi.org/10.1007/978-3-319-10593-2%5C_35 (cit. on pp. 23, 24).

Gordon, Jonathan and Benjamin Van Durme (2013). "Reporting bias and knowledge acquisition". In: *Proceedings of the 2013 workshop on Automated knowledge base construction, AKBC@CIKM 13, San Francisco, California, USA, October 27-28, 2013*, pp. 25–30. URL: https://doi.org/10.1145/2509558.2509563 (cit. on pp. 3, 26, 27, 46, 48, 107).

Goyal, Yash, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh (2017). "Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 6325–6334. URL: https://doi.org/10.1109/CVPR.2017.670 (cit. on p. 40).

Grice, H Paul (1975). "Logic and conversation". In: *1975*, pp. 41–58 (cit. on p. 27).

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. URL: https://doi.org/10.1109/CVPR.2006.100 (cit. on p. 24).

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. URL: https://doi.org/10.1162/0899766042321814 (cit. on p. 23).

Harris, Zellig S (1954). "Distributional structure". In: *Word* 10.2-3, pp. 146–162 (cit. on pp. 14, 17, 43).

He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun (2016). "Deep Residual Learning for Image Recognition". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 770–778. URL: https://doi.org/10.1109/CVPR.2016.90 (cit. on p. 21).

He, Xuming, Richard S. Zemel, and Miguel Á. Carreira-Perpiñán (2004). "Multiscale Conditional Random Fields for Image Labeling". In: *2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2004), with CD-ROM, 27 June - 2 July 2004, Washington, DC, USA*, pp. 695–702. URL: http://doi.ieeecomputersociety.org/10.1109/CVPR.2004.173 (cit. on p. 82).

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. URL: https://doi.org/10.1007/978-3-030-01219-9%5C_47 (cit. on p. 39).

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. URL: http://aclweb.org/anthology/N/N16/N16-1162.pdf (cit. on pp. 16, 17, 69, 70).

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. URL: http://aclweb.org/anthology/D/D14/D14-1032.pdf (cit. on pp. 31, 46, 74).

Hill, Felix, Roi Reichart, and Anna Korhonen (2014b). "Multi-Modal Models for Concrete and Abstract Concept Meaning". In: *TACL* 2, pp. 285–296. URL: https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/393 (cit. on pp. 24, 30, 59).

Hill, Felix, Roi Reichart, and Anna Korhonen (2015). "SimLex-999: Evaluating Semantic Models With (Genuine) Similarity Estimation". In: *Computational Linguistics* 41.4, pp. 665–695. URL: https://doi.org/10.1162/COLI_a_00237 (cit. on p. 32).

Hinton, Geoffrey E (1986). "Learning distributed representations of concepts". In: *Proceedings of the eighth annual conference of the cognitive science society* (cit. on p. 14).

Hinton, Geoffrey E., Oriol Vinyals, and Jeffrey Dean (2015). "Distilling the Knowledge in a Neural Network". In: *CoRR* abs/1503.02531. arXiv: 1503.02531. URL: http://arxiv.org/abs/1503.02531 (cit. on p. 42).

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long Short-Term Memory". In: *Neural Computation* 9.8, pp. 1735–1780. URL: https://doi.org/10.1162/neco.1997.9.8.1735 (cit. on pp. 13, 17).

Hristea, Florentina T. (2011). "Statistical Natural Language Processing". In: *International Encyclopedia of Statistical Science*, pp. 1452–1453. URL: https://doi.org/10.1007/978-3-642-04898-2%5C_82 (cit. on p. 1).

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. URL: http://doi.acm.org/10.1145/1014052.1014073 (cit. on p. 35).

Jastrzebski, Stanislaw, Dzmitry Bahdanau, Seyedarian Hosseini, Michael Noukhovitch, Yoshua Bengio, and Jackie Chi Kit Cheung (2018). "Commonsense mining as knowledge base completion? A study on the impact of novelty". In: *CoRR* abs/1804.09259. arXiv: 1804.09259. URL: http://arxiv.org/abs/1804.09259 (cit. on p. 28).

Jeffreys, Harold (1948). *The theory of probability*. OUP Oxford (cit. on p. 13).

Jiang, Yu-Gang, Jun Yang, Chong-Wah Ngo, and Alexander G. Hauptmann (2010). "Representations of Keypoint-Based Semantic Concept Detection: A Comprehensive Study". In: *IEEE Trans. Multimedia* 12.1, pp. 42–53. URL: https://doi.org/10.1109/TMM.2009.2036235 (cit. on p. 19).

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. URL: https://doi.org/10.1109/CVPR.2017.215 (cit. on p. 40).

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. URL: https://doi.org/10.1109/CVPR.2016.494 (cit. on p. 39).

Jones, Karen Spärck (2004). "A statistical interpretation of term specificity and its application in retrieval". In: *Journal of Documentation* 60.5, pp. 493–502. URL: https://doi.org/10.1108/00220410410560573 (cit. on p. 17).

Jung, Jaewon and Jongyoul Park (2019). "Visual Relationship Detection with Language prior and Softmax". In: *CoRR* abs/1904.07798. arXiv: 1904.07798. URL: http://arxiv.org/abs/1904.07798 (cit. on p. 41).

Kae, Andrew, Gary B. Huang, Carl Doersch, and Erik G. Learned-Miller (2010). "Improving state-of-the-art OCR through high-precision document-specific modeling". In: *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*, pp. 1935–1942. URL: https://doi.org/10.1109/CVPR.2010.5539867 (cit. on p. 2).

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. URL: http://aclweb.org/anthology/P/P14/P14-1062.pdf (cit. on p. 17).

Karpathy, Andrej and Fei-Fei Li (2015). "Deep visual-semantic alignments for generating image descriptions". In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 3128–3137. URL: https://doi.org/10.1109/CVPR.2015.7298932 (cit. on p. 67).

Katz, Slava M. (1987). "Estimation of probabilities from sparse data for the language model component of a speech recognizer". In: *IEEE Trans. Acoustics, Speech, and Signal Processing* 35.3, pp. 400–401. URL: https://doi.org/10.1109/TASSP.1987.1165125 (cit. on p. 11).

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. URL: http://aclweb.org/anthology/D/D14/D14-1005.pdf (cit. on p. 30).

Kiela, Douwe, Luana Bulat, and Stephen Clark (2015a). "Grounding Semantics in Olfactory Perception". 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. 231–236. URL: http://aclweb.org/anthology/P/P15/P15-2038.pdf (cit. on p. 105).

Kiela, Douwe and Stephen Clark (2017). "Learning Neural Audio Embeddings for Grounding Semantics in Auditory Perception". In: *J. Artif. Intell. Res.* 60, pp. 1003–1030. URL: https://doi.org/10.1613/jair.5665 (cit. on p. 105).

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. URL: https://aclanthology.info/papers/N18-1038/n18-1038 (cit. on pp. 34, 64, 65, 69–71, 75).

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. URL: http://aclweb.org/anthology/P/P14/P14-2135.pdf (cit. on p. 46).

Kiela, Douwe, Ivan Vulic, and Stephen Clark (2015b). "Visual Bilingual Lexicon Induction with Transferred ConvNet Features". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015,* *Lisbon, Portugal, September 17-21, 2015*, pp. 148–158. URL: http://aclweb.org/anthology/D/D15/D15-1015.pdf (cit. on p. 37).

Kingma, Diederik P. and Jimmy Ba (2014). "Adam: A Method for Stochastic Optimization". In: *CoRR* abs/1412.6980. arXiv: 1412.6980. URL: http://arxiv.org/abs/1412.6980 (cit. on pp. 71, 94).

Kiros, Ryan, Ruslan Salakhutdinov, and Richard S. Zemel (2014). "Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models". In: *CoRR* abs/1411.2539. arXiv: 1411.2539. URL: http://arxiv.org/abs/1411.2539 (cit. on p. 38).

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. URL: http://papers.nips.cc/paper/5950-skip-thought-vectors (cit. on pp. 16, 17, 66, 69–71, 75).

Klein, Benjamin, Guy Lev, Gil Sadeh, and Lior Wolf (2015). "Associating neural word embeddings with deep image representations using Fisher Vectors". In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 4437–4446. URL: https://doi.org/10.1109/CVPR.2015.7299073 (cit. on p. 24).

Kodirov, Elyor, Tao Xiang, and Shaogang Gong (2017). "Semantic Autoencoder for Zero-Shot Learning". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 4447–4456. URL: https://doi.org/10.1109/CVPR.2017.473 (cit. on p. 80).

Kok, Peter, Michel F Failing, and Floris P de Lange (2014). "Prior expectations evoke stimulus templates in the primary visual cortex". In: *Journal of Cognitive Neuroscience* 26.7, pp. 1546–1554 (cit. on p. 27).

Kordjamshidi, Parisa, Martijn van Otterlo, and Marie-Francine Moens (2010). "Spatial Role Labeling: Task Definition and Annotation Scheme". In: *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2010, 17-23 May 2010, Valletta, Malta*. URL: http://www.lrec-conf.org/proceedings/lrec2010/summaries/846.html (cit. on p. 102).

Kordjamshidi, Parisa, Martijn van Otterlo, and Marie-Francine Moens (2011). "Spatial role labeling: Towards extraction of spatial relations from natural language". In: *TSLP* 8.3, 4:1–4:36. URL: https://doi.org/10.1145/2050104.2050105 (cit. on pp. 101, 106).

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. URL: https://doi.org/10.1109/CVPR.2016.539 (cit. on pp. 30, 36).

Krishna, Ranjay, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei (2017). "Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations". In: *International Journal of Computer Vision* 123.1, pp. 32–73. URL: https://doi.org/10.1007/s11263-016-0981-7 (cit. on pp. 20, 54, 83, 90).

Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton (2012). "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States.* Pp. 1106–1114. URL: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks (cit. on pp. 2, 20, 38).

Lampert, Christoph H., Hannes Nickisch, and Stefan Harmeling (2009). "Learning to detect unseen object classes by between-class attribute transfer". In: *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pp. 951–958. URL: https://doi.org/10.1109/CVPRW.2009.5206594 (cit. on p. 43).

Lampert, Christoph H., Hannes Nickisch, and Stefan Harmeling (2014). "Attribute-Based Classification for Zero-Shot Visual Object Categorization". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 36.3, pp. 453–465. URL: https://doi.org/10.1109/TPAMI.2013.140 (cit. on pp. 43, 81).

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. URL: http://aclweb.org/anthology/N/N16/N16-1030.pdf (cit. on p. 11).

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*. URL: https://openreview.net/forum?id=rkYTTf-AZ (cit. on p. 11).

Lample, Guillaume, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou (2018b). "Word translation without parallel data". In: *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. URL: https://openreview.net/forum?id=H196sainb (cit. on p. 16).

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. URL: http://aclweb.org/anthology/N/N15/N15-1016.pdf (cit. on pp. 31, 46, 55, 57, 64, 69, 75).

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. URL: http://jmlr.org/proceedings/papers/v32/le14.html (cit. on p. 18).

LeCun, Yann, Yoshua Bengio, and Geoffrey E. Hinton (2015). "Deep learning". In: *Nature* 521.7553, pp. 436–444. URL: https://doi.org/10.1038/nature14539 (cit. on p. 10).

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. URL: https://doi.org/10.1162/neco.1989.1.4.541 (cit. on pp. 2, 10, 20).

Lecun, Yann, Léon Bottou, Yoshua Bengio, and Patrick Haffner (1998). "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE*, pp. 2278–2324 (cit. on p. 80).

LeCun, Yann, Sumit Chopra, and Raia Hadsell (2006). "A Tutorial on Energy-Based Learning". In: *Predicting Structured Data* 1, p. 0 (cit. on p. 84).

Lenat, Douglas B., Ramanathan V. Guha, Karen Pittman, Dexter Pratt, and Mary Shepherd (1990). "CYC: Toward Programs With Common Sense". In: *Commun. ACM* 33.8, pp. 30–49. URL: https://doi.org/10.1145/79173.79176 (cit. on p. 28).

Levy, Omer and Yoav Goldberg (2014a). "Dependency-Based Word Embeddings". In: *ACL 2014*. URL: http://aclweb.org/anthology/P/P14/P14-2050.pdf (cit. on p. 50).

Levy, Omer and Yoav Goldberg (2014b). "Linguistic Regularities in Sparse and Explicit Word Representations". In: *Proceedings of the Eighteenth Conference on Computational Natural Language Learning, CoNLL 2014, Baltimore, Maryland, USA, June 26-27, 2014*, pp. 171–180. URL: http://aclweb.org/anthology/W/W14/W14-1618.pdf (cit. on p. 33).

Levy, Omer and Yoav Goldberg (2014c). "Neural Word Embedding as Implicit Matrix Factorization". In: *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pp. 2177–2185. URL: http://papers.nips.cc/paper/5477-neural-word-embedding-as-implicit-matrix-factorization (cit. on pp. 14, 16).

Levy, Omer, Yoav Goldberg, and Ido Dagan (2015). "Improving Distributional Similarity with Lessons Learned from Word Embeddings". In: *TACL* 3, pp. 211–225. URL: https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/570 (cit. on p. 16).

Lin, Tsung-Yi, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick (2014). "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. URL: https://doi.org/10.1007/978-3-319-10602-1%5C_48 (cit. on pp. 20, 69).

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. URL: https://doi.org/10.1109/CVPR.2015.7298917 (cit. on pp. 28, 36, 37).

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. URL: http://arxiv.org/abs/1703.03130 (cit. on p. 17).

Liu, Jingen, Benjamin Kuipers, and Silvio Savarese (2011). "Recognizing human actions by attributes". In: *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011*, pp. 3337–3344. URL: https://doi.org/10.1109/CVPR.2011.5995353 (cit. on p. 43).

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*. URL: https://openreview.net/forum?id=rJvJXZb0W (cit. on p. 17).

Long, Yang, Li Liu, Ling Shao, Fumin Shen, Guiguang Ding, and Jungong Han (2017). "From Zero-Shot Learning to Conventional Supervised Classification: Unseen Visual Data Synthesis". In: *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pp. 6165–6174. URL: https://doi.org/10.1109/CVPR.2017.653 (cit. on pp. 43, 81).

Lowe, David G. (2004). "Distinctive Image Features from Scale-Invariant Keypoints". In: *International Journal of Computer Vision* 60.2, pp. 91–110. URL: https://doi.org/10.1023/B:VISI.0000029664.99615.94 (cit. on pp. 2, 18).

Lu, Cewu, Ranjay Krishna, Michael S. Bernstein, and Fei-Fei Li (2016). "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. URL: https://doi.org/10.1007/978-3-319-46448-0%5C_51 (cit. on pp. 41, 42).

Lu, Yue M. and Minh N. Do (2007). "Multidimensional Directional Filter Banks and Surfacelets". In: *IEEE Trans. Image Processing* 16.4, pp. 918–931. URL: https://doi.org/10.1109/TIP.2007.891785 (cit. on p. 19).

Ludwig, Oswaldo, Xiao Liu, Parisa Kordjamshidi, and Marie-Francine Moens (2016). "Deep Embedding for Spatial Role Labeling". In: *CoRR* abs/1603.08474. URL: http://arxiv.org/abs/1603.08474 (cit. on p. 53).

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. URL: http://www.aclweb.org/anthology/P11-1015 (cit. on p. 11).

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 p. 73).

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. URL: http://papers.nips.cc/paper/5411-a-multi-world-approach-to-question-answering-about-real-world-scenes-based-on-uncertain-input (cit. on p. 39).

Mancini, Massimiliano, José Camacho-Collados, Ignacio Iacobacci, and Roberto Navigli (2017). "Embedding Words and Senses Together via Joint Knowledge-Enhanced Training". In: *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), Vancouver, Canada, August 3-4, 2017*, pp. 100–111. URL: https://doi.org/10.18653/v1/K17-1012 (cit. on p. 61).

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze (2008). *Introduction to information retrieval*. Cambridge University Press (cit. on p. 13).

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. URL: http://aclweb.org/anthology/S/S14/S14-2001.pdf (cit. on p. 35).

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. URL: http://www.lrec-conf.org/proceedings/lrec2014/summaries/363.html (cit. on p. 35).

McRae, Ken, George S Cree, Mark S Seidenberg, and Chris McNorgan (2005). "Semantic feature production norms for a large set of living and nonliving things". In: *Behavior research methods* 37.4, pp. 547–559 (cit. on pp. 31, 32).

Mensink, Thomas, Efstratios Gavves, and Cees G. M. Snoek (2014). "COSTA: Co-Occurrence Statistics for Zero-Shot Classification". In: *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pp. 2441–2448. URL: https://doi.org/10.1109/CVPR.2014.313 (cit. on p. 81).

Mensink, Thomas, Jakob J. Verbeek, Florent Perronnin, and Gabriela Csurka (2012). "Metric Learning for Large Scale Image Classification: Generalizing to New Classes at Near-Zero Cost". In: *Computer Vision - ECCV 2012 - 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part II*, pp. 488–501. URL: https://doi.org/10.1007/978-3-642-33709-3%5C_35 (cit. on pp. 43, 80).

Metzler, Donald (2008). "Beyond bags of words: effectively modeling dependence and features in information retrieval". In: *SIGIR Forum* 42.1, p. 77. URL: https://doi.org/10.1145/1394251.1394271 (cit. on p. 12).

Mikolov, Tomas, Anoop Deoras, Stefan Kombrink, Lukás Burget, and Jan Cernocký (2011). "Empirical Evaluation and Combination of Advanced Language Modeling Techniques". In: *INTERSPEECH 2011, 12th Annual Conference of the International Speech Communication Association, Florence, Italy, August 27-31, 2011*, pp. 605–608. URL: http://www.isca-speech.org/archive/interspeech%5C_2011/i11%5C_0605.html (cit. on p. 14).

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean (2013). "Distributed Representations of Words and Phrases and their Compositionality". 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. 3111–3119. URL: http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality (cit. on pp. 2, 14, 15, 30, 33, 43, 49, 54, 82, 84, 94).

Miller, George A. (1995). "WordNet: A Lexical Database for English". In: *Commun. ACM* 38.11, pp. 39–41. URL: http://doi.acm.org/10.1145/219717.219748 (cit. on p. 29).

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. URL: https://doi.org/10.1007/978-3-319-58068-5%5C_21 (cit. on p. 11).

Mostafazadeh, Nasrin, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios P. Spithourakis, and Lucy Vanderwende (2017). "Image-Grounded Conversations: Multimodal Context for Natural Question and Response Generation". In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1,* *2017 - Volume 1: Long Papers*, pp. 462–472. URL: https://aclanthology.info/papers/I17-1047/i17-1047 (cit. on p. 40).

Nadas, Arthur (1984). "Estimation of probabilities in the language model of the IBM speech recognition system". In: *IEEE Transactions on Acoustics, Speech, and Signal Processing* 32.4, pp. 859–861 (cit. on p. 13).

Nelson, Douglas L, Cathy L McEvoy, and Thomas A Schreiber (2004). "The University of South Florida free association, rhyme, and word fragment norms". In: *Behavior Research Methods, Instruments, & Computers* (cit. on pp. 33, 73).

Nenov, Valeriy I and Michael G Dyer (1988). "DETE: Connectionist/symbolic model of visual and verbal association". In: *Proceedings of The connectionnist models summer school 1988* (cit. on p. 29).

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. URL: https://doi.org/10.3233/AIC-150698 (cit. on p. 11).

Norman, Donald A (1972). "Memory, knowledge, and the answering of questions." In: (cit. on pp. 16, 64).

Norouzi, Mohammad, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg Corrado, and Jeffrey Dean (2013). "Zero-Shot Learning by Convex Combination of Semantic Embeddings". In: *CoRR* abs/1312.5650. arXiv: 1312.5650. URL: http://arxiv.org/abs/1312.5650 (cit. on p. 43).

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. URL: https://aclanthology.info/papers/D17-1238/d17-1238 (cit. on p. 39).

Palatucci, Mark, Dean Pomerleau, Geoffrey E. Hinton, and Tom M. Mitchell (2009). "Zero-shot Learning with Semantic Output Codes". In: *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009. Proceedings of a meeting held 7-10 December 2009, Vancouver, British Columbia, Canada.* Pp. 1410–1418. URL: http://papers.nips.cc/paper/3650-zero-shot-learning-with-semantic-output-codes (cit. on p. 43).

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. URL: http://aclweb.org/anthology/P/P04/P04-1035.pdf (cit. on p. 35).

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. URL: http://aclweb.org/anthology/P/P05/P05-1015.pdf (cit. on p. 35).

Pang, Bo and Lillian Lee (2007). "Opinion Mining and Sentiment Analysis". In: *Foundations and Trends in Information Retrieval* 2.1-2, pp. 1–135. URL: https://doi.org/10.1561/1500000011 (cit. on pp. 1, 11).

Parameswaran, Vasu and Rama Chellappa (2004). "View Independent Human Body Pose Estimation from a Single Perspective Image". In: *2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2004), with CD-ROM, 27 June - 2 July 2004, Washington, DC, USA*, pp. 16–22. URL: http://doi.ieeecomputersociety.org/10.1109/CVPR.2004.264 (cit. on p. 2).

Parikh, Devi and Kristen Grauman (2011). "Interactively building a discriminative vocabulary of nameable attributes". In: *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011*, pp. 1681–1688. URL: https://doi.org/10.1109/CVPR.2011.5995451 (cit. on p. 43).

Pennington, Jeffrey, Richard Socher, and Christopher D. Manning (2014). "Glove: Global Vectors for Word Representation". 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. 1532–1543. URL: http://aclweb.org/anthology/D/D14/D14-1162.pdf (cit. on pp. 14, 15, 30, 43).

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. URL: https://aclanthology.info/papers/N18-1202/n18-1202 (cit. on pp. 16, 43).

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. URL: http://www.lrec-conf.org/proceedings/lrec2012/summaries/274.html (cit. on pp. 1, 11).

Pezeshkpour, Pouya, Liyan Chen, and Sameer Singh (2018). "Embedding Multimodal Relational Data for Knowledge Base Completion". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pp. 3208–3218. URL: https://aclanthology.info/papers/D18-1359/d18-1359 (cit. on p. 61).

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. URL: https://doi.org/10.1109/ICCV.2015.303 (cit. on p. 73).

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. URL: https://doi.org/10.1007/s11263-016-0965-7 (cit. on p. 24).

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. URL: http://aclweb.org/anthology/S/S16/S16-1002.pdf (cit. on p. 11).

Qiao, Ruizhi, Lingqiao Liu, Chunhua Shen, and Anton van den Hengel (2016). "Less is More: Zero-Shot Learning from Online Textual Documents with Noise Suppression". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 2249–2257. URL: https://doi.org/10.1109/CVPR.2016.247 (cit. on p. 43).

Qiu, Guoping (2002). "Indexing chromatic and achromatic patterns for content-based colour image retrieval". In: *Pattern Recognition* 35.8, pp. 1675–1686. URL: https://doi.org/10.1016/S0031-3203(01)00162-5 (cit. on p. 18).

Rabinovich, Andrew, Andrea Vedaldi, Carolina Galleguillos, Eric Wiewiora, and Serge J. Belongie (2007). "Objects in Context". In: *IEEE 11th International Conference on Computer Vision, ICCV 2007, Rio de Janeiro, Brazil, October 14-20, 2007*, pp. 1–8. URL: https://doi.org/10.1109/ICCV.2007.4408986 (cit. on p. 82).

Radford, Alec, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever (2019). "Language Models are Unsupervised Multitask Learners". In: (cit. on p. 14).

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. URL: http://papers.nips.cc/paper/7427-overcoming-language-priors-in-visual-question-answering-with-adversarial-regularization (cit. on p. 40).

Ramanathan, Vignesh, Armand Joulin, Percy Liang, and Fei-Fei Li (2014). "Linking People in Videos with "Their" Names Using Coreference Resolution". In:

Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I, pp. 95–110. URL: https://doi.org/10.1007/978-3-319-10590-1%5C_7 (cit. on p. 107).

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*. URL: http://arxiv.org/abs/1511.06732 (cit. on p. 39).

Ratan, Aparna Lakshmi, W. Eric L. Grimson, and William M. Wells III (1998). "Object Detection and Localization by Dynamic Template Warping". In: *1998 Conference on Computer Vision and Pattern Recognition (CVPR '98), June 23-25, 1998, Santa Barbara, CA, USA*, pp. 634–640. URL: https://doi.org/10.1109/CVPR.1998.698671 (cit. on p. 2).

Real, Esteban, Alok Aggarwal, Yanping Huang, and Quoc V. Le (2018). "Regularized Evolution for Image Classifier Architecture Search". In: *CoRR* abs/1802.01548. arXiv: 1802.01548. URL: http://arxiv.org/abs/1802.01548 (cit. on pp. 42, 80).

Rehurek, Radim and Petr Sojka (n.d.). "Software Framework for Topic Modelling with Large Corpora". In: *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* (cit. on p. 54).

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. URL: https://doi.org/10.1109/TPAMI.2016.2577031 (cit. on pp. 101, 106).

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. URL: https://doi.org/10.1109/CVPR.2017.128 (cit. on p. 39).

Rohrbach, Anna, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele (2016). "Grounding of Textual Phrases in Images by Reconstruction". In: *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*, pp. 817–834. URL: https://doi.org/10.1007/978-3-319-46448-0%5C_49 (cit. on p. 22).

Roller, Stephen and Sabine Schulte im Walde (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, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pp. 1146–1157. URL: https://aclanthology.info/papers/D13-1115/d13-1115 (cit. on pp. 30, 47, 48).

Romera-Paredes, Bernardino and Philip H. S. Torr (2015). "An embarrassingly simple approach to zero-shot learning". In: *Proceedings of the 32nd International*

Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015, pp. 2152–2161. URL: http://jmlr.org/proceedings/papers/v37/romera-paredes15.html (cit. on p. 43).

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. URL: http://aclweb.org/anthology/D/D15/D15-1044.pdf (cit. on p. 11).

Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei (2015). "ImageNet Large Scale Visual Recognition Challenge". In: *International Journal of Computer Vision (IJCV)* 115.3, pp. 211–252 (cit. on p. 38).

Sadeghi, Fereshteh, Santosh Kumar Divvala, and Ali Farhadi (2015). "VisKE: Visual knowledge extraction and question answering by visual verification of relation phrases". In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 1456–1464. URL: https://doi.org/10.1109/CVPR.2015.7298752 (cit. on pp. 28, 41).

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. URL: https://doi.org/10.1109/CVPR.2011.5995711 (cit. on p. 41).

Sagara, Tsukasa and Masafumi Hagiwara (2014). "Natural language neural network and its application to question-answering system". In: *Neurocomputing* 142, pp. 201–208. URL: https://doi.org/10.1016/j.neucom.2014.04.048 (cit. on p. 16).

Salton, Gerard and Michael McGill (1984). *Introduction to Modern Information Retrieval*. McGraw-Hill Book Company (cit. on p. 17).

Salton, Gerard, A. Wong, and Chung-Shu Yang (1975). "A Vector Space Model for Automatic Indexing". In: *Commun. ACM* 18.11, pp. 613–620. URL: https://doi.org/10.1145/361219.361220 (cit. on p. 1).

Schäfer, Anton Maximilian, Steffen Udluft, and Hans-Georg Zimmermann (2006). "Learning Long Term Dependencies with Recurrent Neural Networks". In: *Artificial Neural Networks - ICANN 2006, 16th International Conference, Athens, Greece, September 10-14, 2006. Proceedings, Part I*, pp. 71–80. URL: https://doi.org/10.1007/11840817%5C_8 (cit. on p. 14).

Schakel, Adriaan M. J. and Benjamin J. Wilson (2015a). "Measuring Word Significance using Distributed Representations of Words". In: *CoRR* abs/1508.02297. arXiv: 1508.02297. URL: http://arxiv.org/abs/1508.02297 (cit. on p. 87).

Schakel, Adriaan M. J. and Benjamin J. Wilson (2015b). "Measuring Word Significance using Distributed Representations of Words". In: *CoRR* abs/1508.02297. arXiv: 1508.02297. URL: http://arxiv.org/abs/1508.02297 (cit. on p. 95).

Schroff, Florian, Dmitry Kalenichenko, and James Philbin (2015). "FaceNet: A unified embedding for face recognition and clustering". In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pp. 815–823. URL: https://doi.org/10.1109/CVPR.2015.7298682 (cit. on p. 24).

Schütze, Hinrich (1992). "Word Space". In: *Advances in Neural Information Processing Systems 5, [NIPS Conference, Denver, Colorado, USA, November 30 - December 3, 1992]*, pp. 895–902. URL: http://papers.nips.cc/paper/603-word-space (cit. on p. 14).

Schwenk, Holger (2007). "Continuous space language models". In: *Computer Speech & Language* 21.3, pp. 492–518. URL: https://doi.org/10.1016/j.csl.2006.09.003 (cit. on p. 14).

Schwenk, Holger, Loıc Barrault, Alexis Conneau, and Yann LeCun (2017). "Very Deep Convolutional Networks for Text Classification". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers*, pp. 1107–1116. URL: https://aclanthology.info/papers/E17-1104/e17-1104 (cit. on p. 14).

Sennrich, Rico, Barry Haddow, and Alexandra Birch (2016). "Neural Machine Translation of Rare Words with Subword Units". 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*. URL: http://aclweb.org/anthology/P/P16/P16-1162.pdf (cit. on p. 14).

Shalaby, Walid, Wlodek Zadrozny, and Hongxia Jin (2018). "Beyond Word Embeddings: Learning Entity and Concept Representations from Large Scale Knowledge Bases". In: *CoRR* abs/1801.00388. arXiv: 1801.00388. URL: http://arxiv.org/abs/1801.00388 (cit. on p. 105).

Shannon, Claude E (1951). "Prediction and entropy of printed English". In: *Bell system technical journal* 30.1, pp. 50–64 (cit. on p. 13).

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 2012, July 12-14, 2012, Jeju Island, Korea*, pp. 1423–1433. URL: http://www.aclweb.org/anthology/D12-1130 (cit. on pp. 23, 30).

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. URL: http://aclweb.org/anthology/P/P14/P14-1068.pdf (cit. on pp. 32, 65).

Simonyan, Karen and Andrew Zisserman (2015). "Very Deep Convolutional Networks for Large-Scale Image Recognition". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015,* *Conference Track Proceedings*. URL: http://arxiv.org/abs/1409.1556 (cit. on pp. 20, 21).

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*. URL: https://openreview.net/forum?id=r1Aab85gg (cit. on p. 16).

Socher, Richard, Andrej Karpathy, Quoc V. Le, Christopher D. Manning, and Andrew Y. Ng (2014). "Grounded Compositional Semantics for Finding and Describing Images with Sentences". In: *TACL* 2, pp. 207–218. URL: https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/325 (cit. on p. 24).

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. URL: https://aclanthology.info/papers/D13-1170/d13-1170 (cit. on pp. 17, 35).

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*, pp. 196–205. URL: http://aclweb.org/anthology/N/N15/N15-1020.pdf (cit. on p. 11).

Speer, Robyn, Joshua Chin, and Catherine Havasi (2017). "ConceptNet 5.5: An Open Multilingual Graph of General Knowledge". In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.* Pp. 4444–4451. URL: http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14972 (cit. on pp. 12, 28).

Srivastava, Nitish, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov (2014). "Dropout: a simple way to prevent neural networks from overfitting". In: *Journal of Machine Learning Research* 15.1, pp. 1929–1958. URL: http://dl.acm.org/citation.cfm?id=2670313 (cit. on p. 20).

Szegedy, Christian, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna (2016). "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. URL: https://doi.org/10.1109/CVPR.2016.308 (cit. on pp. 21, 57, 71, 94).

Tian, Fei, Bin Gao, Enhong Chen, and Tie-Yan Liu (2016). "Learning Better Word Embedding by Asymmetric Low-Rank Projection of Knowledge Graph". In: *J.*

Comput. Sci. Technol. 31.3. URL: https://doi.org/10.1007/s11390-016-1651-5 (cit. on p. 16).

Tissier, Julien, Christophe Gravier, and Amaury Habrard (2017). "Dict2vec : Learning Word Embeddings using Lexical Dictionaries". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pp. 254–263. URL: https://aclanthology.info/papers/D17-1024/d17-1024 (cit. on p. 105).

Torralba, Antonio (2003). "Contextual Priming for Object Detection". In: *International Journal of Computer Vision* 53.2, pp. 169–191. URL: https://doi.org/10.1023/A:1023052124951 (cit. on p. 82).

Torralba, Antonio, Kevin P. Murphy, and William T. Freeman (2010). "Using the forest to see the trees: exploiting context for visual object detection and localization". In: *Commun. ACM* 53.3, pp. 107–114. URL: http://doi.acm.org/10.1145/1666420.1666446 (cit. on pp. 82, 94).

Toutanova, Kristina, Danqi Chen, Patrick Pantel, Hoifung Poon, Pallavi Choudhury, and Michael Gamon (2015). "Representing Text for Joint Embedding of Text and Knowledge Bases". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pp. 1499–1509. URL: http://aclweb.org/anthology/D/D15/D15-1174.pdf (cit. on p. 61).

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. URL: https://doi.org/10.1007/s11263-013-0620-5 (cit. on pp. 101, 106).

Vanderwende, Lucy (2005). "Volunteers Created the Web". 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. 84–90. URL: http://www.aaai.org/Library/Symposia/Spring/2005/ss05-03-013.php (cit. on p. 28).

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. 6000–6010. URL: http://papers.nips.cc/paper/7181-attention-is-all-you-need (cit. on p. 14).

Vedantam, Ramakrishna, Xiao Lin, Tanmay Batra, C. Lawrence Zitnick, and Devi Parikh (2015). "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. URL: https://doi.org/10.1109/ICCV.2015.292 (cit. on p. 28).

Vilnis, Luke and Andrew McCallum (2015). "Word Representations via Gaussian Embedding". In: *3rd International Conference on Learning Representations, ICLR*

2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings. URL: http://arxiv.org/abs/1412.6623 (cit. on p. 16).

Vinyals, Oriol and Quoc V. Le (2015). "A Neural Conversational Model". In: *CoRR* abs/1506.05869. arXiv: 1506.05869. URL: http://arxiv.org/abs/1506.05869 (cit. on p. 11).

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. URL: https://doi.org/10.1109/CVPR.2017.475 (cit. on p. 40).

Vukotic, Vedran, Christian Raymond, and Guillaume Gravier (2015). "Is it time to Switch to word embedding and recurrent neural networks for spoken language understanding?" In: *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, pp. 130–134. URL: http://www.isca-speech.org/archive/interspeech%5C_2015/i15%5C_0130.html (cit. on p. 11).

Vukotic, Vedran, Christian Raymond, and Guillaume Gravier (2018). "A Cross-modal Approach to Multimodal Fusion in Video Hyperlinking". In: *IEEE MultiMedia* 25.2, pp. 11–23. URL: https://doi.org/10.1109/MMUL.2018.023121161 (cit. on p. 23).

Vulic, Ivan, Douwe Kiela, Stephen Clark, and Marie-Francine Moens (2016). "Multi-Modal Representations for Improved Bilingual Lexicon Learning". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 2: Short Papers*. URL: http://aclweb.org/anthology/P/P16/P16-2031.pdf (cit. on p. 37).

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. URL: https://doi.org/10.1145/2766462.2767752 (cit. on p. 17).

Wah, Catherine, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie (2011). "The caltech-ucsd birds-200-2011 dataset". In: (cit. on p. 90).

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. URL: http://arxiv.org/abs/1903.02588 (cit. on p. 64).

Weinberger, Kilian Q. and Lawrence K. Saul (2009). "Distance Metric Learning for Large Margin Nearest Neighbor Classification". In: *Journal of Machine Learning Research* 10, pp. 207–244. URL: https://dl.acm.org/citation.cfm?id=1577078 (cit. on p. 25).

Weston, Jason, Samy Bengio, and Nicolas Usunier (2010). "Large scale image annotation: learning to rank with joint word-image embeddings". In: *Machine*

Learning 81.1, pp. 21–35. URL: https://doi.org/10.1007/s10994-010-5198-3 (cit. on p. 23).

Weston, Jason, Samy Bengio, and Nicolas Usunier (2011). "WSABIE: Scaling Up to Large Vocabulary Image Annotation". In: *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011*, pp. 2764–2770. URL: https://doi.org/10.5591/978-1-57735-516-8/IJCAI11-460 (cit. on p. 87).

Weston, Jason, Antoine Bordes, Oksana Yakhnenko, and Nicolas Usunier (2013). "Connecting Language and Knowledge Bases with Embedding Models for Relation Extraction". 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. 1366–1371. URL: http://aclweb.org/anthology/D/D13/D13-1136.pdf (cit. on pp. 61, 105).

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*. URL: http://arxiv.org/abs/1410.3916 (cit. on p. 11).

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 p. 35).

Wolf, Lior and Stanley M. Bileschi (2006). "A Critical View of Context". In: *International Journal of Computer Vision* 69.2, pp. 251–261. URL: https://doi.org/10.1007/s11263-006-7538-0 (cit. on p. 82).

Wu, Hui, Michele Merler, Rosario Uceda-Sosa, and John R. Smith (2016). "Learning to Make Better Mistakes: Semantics-aware Visual Food Recognition". In: *Proceedings of the 2016 ACM Conference on Multimedia Conference, MM 2016, Amsterdam, The Netherlands, October 15-19, 2016*, pp. 172–176. URL: https://doi.org/10.1145/2964284.2967205 (cit. on p. 106).

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. URL: https://doi.org/10.1145/3077136.3080678 (cit. on p. 23).

Xian, Yongqin, Zeynep Akata, Gaurav Sharma, Quynh N. Nguyen, Matthias Hein, and Bernt Schiele (2016). "Latent Embeddings for Zero-Shot Classification". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 69–77. URL: https://doi.org/10.1109/CVPR.2016.15 (cit. on p. 43).

Xiao, Jianxiong, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba (2010). "SUN database: Large-scale scene recognition from abbey to zoo". In:

The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010, pp. 3485–3492. URL: https://doi.org/10.1109/CVPR.2010.5539970 (cit. on p. 90).

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. URL: http://papers.nips.cc/paper/2164-distance-metric-learning-with-application-to-clustering-with-side-information (cit. on p. 24).

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. URL: http://jmlr.org/proceedings/papers/v37/xuc15.html (cit. on p. 39).

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. URL: http://aclweb.org/anthology/N/N16/N16-1023.pdf (cit. on p. 28).

Yu, Fisher and Vladlen Koltun (2016). “Multi-Scale Context Aggregation by Dilated Convolutions”. In: *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*. URL: http://arxiv.org/abs/1511.07122 (cit. on p. 81).

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. URL: https://doi.org/10.1109/ICCV.2017.121 (cit. on pp. 5, 41, 42).

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. URL: https://doi.org/10.1109/ICCV.2017.202 (cit. on pp. 22, 40).

Zablocki, Eloi, Patrick Bordes, Laure Soulier, Benjamin Piwowarski, and Patrick Gallinari (2019). “Context-Aware Zero-Shot Learning for Object Recognition”. In: *ICML 2019* (cit. on pp. 7, 80).

Zablocki, Eloi, Benjamin Piwowarski, Laure Soulier, and Patrick Gallinari (2018). “Learning Multi-Modal Word Representation Grounded in Visual Context”. In: *AAAI 2018* (cit. on pp. 6, 46).

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. URL: https://doi.org/10.1007/978-3-319-10590-1%5C_53 (cit. on pp. 2, 19).

Zhang, Xiaofan, Feng Zhou, Yuanqing Lin, and Shaoting Zhang (2016). "Embedding Label Structures for Fine-Grained Feature Representation". In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 1114–1123. URL: https://doi.org/10.1109/CVPR.2016.126 (cit. on p. 106).

Zhao, Bo, Bo Chang, Zequn Jie, and Leonid Sigal (2018). "Modular Generative Adversarial Networks". In: *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XIV*, pp. 157–173. URL: https://doi.org/10.1007/978-3-030-01264-9%5C_10 (cit. on p. 90).

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. URL: https://doi.org/10.1109/ICCV.2015.11 (cit. on p. 24).

Zitnick, C. Lawrence and Piotr Dollár (2014). "Edge Boxes: Locating Object Proposals from Edges". In: *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, pp. 391–405. URL: https://doi.org/10.1007/978-3-319-10602-1%5C_26 (cit. on pp. 101, 106).

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. URL: https://doi.org/10.1109/CVPR.2013.387 (cit. on p. 28).

Zoph, Barret, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le (2017). "Learning Transferable Architectures for Scalable Image Recognition". In: *CoRR* abs/1707.07012. arXiv: 1707.07012. URL: http://arxiv.org/abs/1707.07012 (cit. on pp. 42, 80).