Incorporating Visual Semantics into Sentence Representations within a Grounded Space

Patrick Bordes, Éloi Zablocki, Laure Soulier, Benjamin Piwowarski, Patrick Gallinari

HAL Id: hal-02351003

https://hal.sorbonne-universite.fr/hal-02351003v1

Submitted on 5 Oct 2021

Incorporating Visual Semantics into Sentence Representations within a Grounded Space

Patrick Bordes[*1] and **Éloi Zablocki**[*1] and **Laure Soulier**[1]
and **Benjamin Piwowarski**[1] and **Patrick Gallinari**[1,2]

[1] Sorbonne Université, CNRS, LIP6, F-75005 Paris, France
[2] Criteo AI Lab, Paris, France

`firstname.name@lip6.fr`

Abstract

Language grounding is an active field aiming at enriching textual representations with visual information. Generally, textual and visual elements are embedded in the same representation space, which implicitly assumes a one-to-one correspondence between modalities. This hypothesis does not hold when representing words, and becomes problematic when used to learn sentence representations — the focus of this paper — as a visual scene can be described by a wide variety of sentences. To overcome this limitation, we propose to transfer visual information to textual representations by learning an intermediate representation space: the grounded space. We further propose two new complementary objectives ensuring that (1) sentences associated with the same visual content are close in the grounded space and (2) similarities between related elements are preserved across modalities. We show that this model outperforms the previous state-of-the-art on classification and semantic relatedness tasks.

1 Introduction

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

sentences describe relationships between objects, and thus convey complex and high-level knowledge better than individual words (Norman, 1972).

Relying only on text can lead to biased representations and unrealistic predictions such as "*the sky is green*" (Baroni, 2016). Besides, it has been shown that human understanding of language is *grounded* in physical reality and perceptual experience (Fincher-Kiefer, 2001). To overcome this limitation, an emerging approach is to *ground* language in the visual world: this consists in leveraging visual information, usually from images, to enrich textual representations.[*]

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

To overcome this issue, we propose an alter-

---

[*] Equal contribution.

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

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

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

2   Related work

Over the last years, several approaches have been proposed to learn semantic representations for sentences. This includes supervised and task-specific techniques with recursive networks (Socher et al., 2013), convolutional networks (Kalchbrenner et al., 2014) or self-attentive networks (Lin et al., 2017; Conneau et al., 2017), but also unsupervised methods producing universal representations given large text corpora. Examples of the latter include models such as FastSent (Hill et al., 2016), QuickThought (Logeswaran and Lee, 2018), Word Information Series (Arroyo-Fernández et al., 2019), Universal Sentence Encoder (Cer et al., 2018), or SkipThought (Kiros et al., 2015), where a sentence is encoded with a Gated Recurrent Unit (GRU), and two GRU decoders are trained to reconstruct the adjacent sentences in a dataset of ordered sentences.

To model the way language conveys meaning, traditional approaches consider language as a purely symbolic system based on words and syntactic rules (Chomsky, 1980; Burgess and Lund,

1997). However, W. Barsalou (1999); Fincher-Kiefer (2001) insist on the intuition that language has to be grounded in physical reality and perceptual experience. The importance of language grounding is underlined in Gordon and Van Durme (2013), where an important bias is reported: the frequency at which objects, relations, or events occur in natural language is significantly different from their real-world frequency (e.g., in texts, people are *murdered* four times more than they *breathe*). Thus, leveraging visual resources, in addition to textual resources, is a promising way to acquire commonsense knowledge (Lin and Parikh, 2015; Yatskar et al., 2016) and to cope with the bias between text and reality.

This intuition has motivated several works for learning visually grounded representations for words, using images — or abstract scenes (Kottur et al., 2016). Two lines of work can be distinguished. First, *sequential* techniques combine textual and visual representations that were learned separately (Bruni et al., 2014; Silberer and Lapata, 2014; Collell et al., 2017). Second, *joint* methods learn a multimodal representation from multiple sources simultaneously. The advantage is that visual information associated with concrete words can be transferred to more abstract ones, which usually have no associated visual data (Hill et al., 2014; Lazaridou et al., 2015b). Closer to our contribution, some approaches learn grounded word embeddings by building upon the skip-gram objective (Mikolov et al., 2013) and enforcing word vectors to be close to their corresponding visual features (Lazaridou et al., 2015b) or their visual context (Zablocki et al., 2018) in a multimodal representation space. These approaches learn word representations while we specifically target sentences. This task is more challenging since sentences are inherently different than words due to their sequential and compositional nature. Moreover, a great number of different sentences can be generated for the same image.

Finally, some works learn sentence representations by aligning visual data to sentences within 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 and the other one is trained to predict the visual features associated to a sentence. The model of Kiela et al. (2018) is close to IMAGINET and additionally
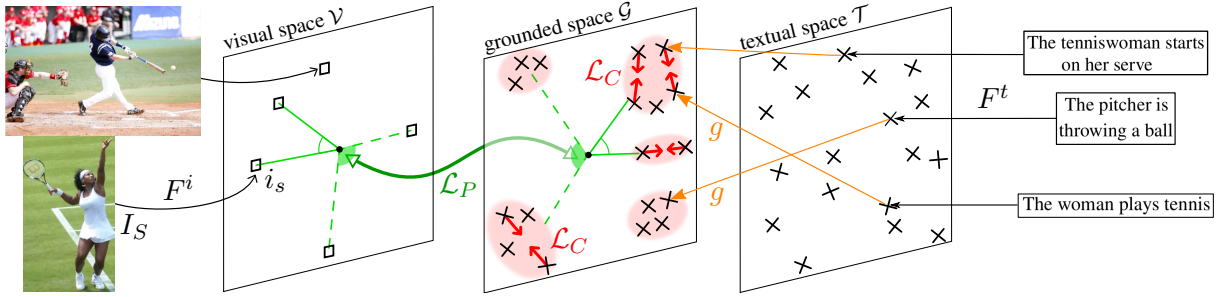
Figure 1: Model overview. Red circles indicate visual clusters. Red arrows represent the gradient of the cluster loss, which gathers visually equivalent sentences — the contrastive term in loss $\mathcal{L}_C$ is not represented. The green arrow and angles illustrate the perceptual loss, ensuring that cosine similarities correlate across modalities. The origin is at the center of each space.

hypothesizes that associated captions ground the meaning of a sentence. Both of these works learn a projection of the sentence representation to the corresponding image, which we argue is problematic as it over-constrains the textual space and can degrade textual representations. Indeed, Collell and Moens (2018) empirically demonstrated that when a cross-modal mapping is learned, the projection of the source modality does not resemble the target modality, in the sense of nearest neighbor comparison. This suggests that cross-modal projections are not appropriate to incorporate visual semantics in text representations.

3 Incorporating visual semantics within an intermediate grounded space

3.1 Model overview

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

$$\mathcal{L}(\theta^t, \theta^i) = \mathcal{L}_\mathcal{T}(\theta^t) + \mathcal{L}_\mathcal{G}(\theta^t, \theta^i) \quad (1)$$

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

In what follows, we focus on the modeling of the grounding objective $\mathcal{L}_\mathcal{G}$, learned on a captioning corpus, where each image is associated with several captions. Grounding approaches generally leverage visual information by embedding textual and visual elements within the same multimodal space (Silberer and Lapata, 2014; Kiela et al., 2018). However, it is not satisfying since texts and images are forced to be in one-to-one correspondence. Moreover, a caption can (1) have a wide variety of paraphrases and related sentences describing the same scene (e.g., *the kitten is devouring a mouse* versus *a cat eating a mouse*), (2) be visually ambiguous (e.g., *a cat is eating* can be associated with many different images, depending on the visual scene/context), or (3) carry non-visual information (e.g., *cats often think about their meals*). Usual grounding objectives, that embed sentences in the visual space, can discard non-visual information (3) through the projection function. They can handle (1) by projecting related sentences to the same location in the visual space. However, they are over-sensitive to visual ambiguity (2), because ambiguous sentences should be projected to different locations of the visual space, which is not possible with current grounding models.

To overcome this lack of flexibility, we propose the following approach, illustrated in Figure 1. To cope with (1), sentences associated with the same image should be close — we call this *cluster information*. To cope with (2), we want to avoid projecting sentences to a particular point of the visual space: instead, we require that the similarity between two images in the visual space (which is linked to the context discrepancy) should be close to the similarity between their associated sentences in the textual space. We call this *perceptual information*. Finally, as we want to preserve non-visual information in sentence representations (3), we make use of an intermediate space,

698

called *grounded space*, that allows textual representations to benefit from visual properties without degrading the semantics brought by the textual objective $\mathcal{L}_\mathcal{T}$.

3.2 Grounding space and objectives

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

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

Cluster information (C_g) The cluster information leverages the fact that two sentences describe, or not, the same underlying reality. In other words, the goal is to measure if two sentences are *visually equivalent* (assumption (1) in Section 3.1) without considering the content of related images. For convenience, two sentences are said to be *visually equivalent* (resp. *visually different*) if they are associated with the same image (resp. different images), i.e. if they describe the same (resp. different) underlying reality. We call *cluster* a set of visually equivalent sentences. For instance, in Figure 1, sentences *The tenniswoman starts on her serve* and *The woman plays tennis* are visually equivalent and belong to the same cluster.

Our hypothesis is that *the similarity between visually equivalent sentences* (s, s^+) *should be higher than visually different sentences* (s, s^-). We translate this hypothesis into the constraint in the grounded space: $\cos(g(s), g(s^+)) \leq \cos(g(s), g(s^-))$. Following (Karpathy and Li, 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 1) resulting

in the cluster loss $\mathcal{L}_\mathcal{C}$:

$$\mathcal{L}_C = \sum_{(s,s^+,s^-)} \left\lfloor \gamma - \cos(g(s), g(s^+)) + \cos(g(s), g(s^-)) \right\rfloor_+ \tag{2}$$

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

Perceptual information (P_g) The cluster hypothesis alone ignores the structure of the visual space and only uses the visual modality as a proxy to assess if two sentences are visually equivalent or different. Moreover, the ranking loss $\mathcal{L}_\mathcal{C}$ simply drives apart visually different sentences in the representation space, which can be a problem when two images have a closely related content. For instance, the baseball and tennis images in Figure 1 may be different, but they are both sports images, and thus their corresponding sentences should be somehow close in the grounded space. Finally, it supposes that we have a dataset of images associated with several captions.

To cope with these limitations, we consider the structure of the visual space and use the content of images. The intuition is that the structure of the textual space should be modeled on the structure of the visual one to extract visual semantics. We choose to preserve *similarities* between related elements across spaces (cf. green elements in Figure 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 $\mathcal{L}_\mathcal{P}$:

$$\mathcal{L}_\mathcal{P} = -\rho(\{sim_{k_1,k_2}^{\text{text}}\}, \{sim_{k_1,k_2}^{\text{im}}\}) \tag{3}$$

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

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

$$\mathcal{L}_\mathcal{G}(\theta^t, \theta^i) = \alpha_C \mathcal{L}_C(\theta^t, \theta^i) + \alpha_P \mathcal{L}_P(\theta^t, \theta^i) \tag{4}$$

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

4 Evaluation protocol

4.1 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 MS COCO (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 number of sentences in the training set of COCO (590K sentences) only represents 0.8% of the sentence data in BookCorpus, which is negligible, and the additional textual training data cannot account for performance discrepancies between textual and grounded models.

4.2 Baselines and Scenarios

In the experiments, we focus on one of the most established sentence models: SkipThought (noted **T**) as the textual baseline: the parameters of the sentence embedding model are obtained by minimizing $\mathcal{L}_{\mathcal{T}}$. Then, we derive several baselines and scenarios based on **T**, each representing a different approach of grounding. Since our focus is to study the impact of grounding on sentence representations, all baselines and scenarios share the same representation dimension $d_t = 2048$ and are trained on the same datasets (cf. sect. 4.1). We also report a textual model of dimension $\frac{d_t}{2}$ that we call $\mathbf{T}_{1024}$, to compare with the GroundSent model of (Kiela et al., 2018).

Model Scenarios. We test variants of our grounding model presented in sect. 3, all based on **T**: $\mathbf{T} + \mathbf{C}_g$, $\mathbf{T} + \mathbf{P}_g$, $\mathbf{T} + \mathbf{C}_g + \mathbf{P}_g$, where $\mathbf{C}_g$ (resp. $\mathbf{P}_g$) represents the loss $\mathcal{L}_C$ (resp. $\mathcal{L}_P$). We also consider scenarios where g equals the identity function (no grounded space), which we note $\mathbf{C}_{id}$, $\mathbf{P}_{id}$, $\mathbf{C}_{id} + \mathbf{P}_{id}$, etc. Finally, we also performed preliminary analysis learning only from the visual modality: $\mathbf{C}_{g/id}$, $\mathbf{P}_{g/id}$, $\mathbf{C}_{g/id} + \mathbf{P}_{g/id}$.

Baselines. We adapt two classical multimodal word embedding models for sentences. Accordingly, models from the two existing model families are considered:

Cross-modal Projection (**CM**): Inspired by Lazaridou et al. (2015b), this baseline learns to project sentences in the visual space using a max-margin loss:

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

where f is a MLP, γ' a fixed margin and i^- a non-matching image. Similarly to our scenarios, the sentence encoder is initialized with **T**.

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

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

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

4.3 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:

Semantic relatedness. We use two semantic similarity benchmarks: STS (Cer et al., 2017) and SICK (Marelli et al., 2014a), which consist of

	Structural measures				Semantic relatedness				
Model	mNNO	ρ_{vis}	C_{inter}	C_{intra}	STS/All	STS/Cap	STS/News	STS/Forum	SICK
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**
CM (vis.)	27.1	19.2	1.5	85.8	56	78	40	34	55
$\mathbf{P}_g$	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 1: Intrinsic evaluations carried out on the grounded space for models with $g =$ MLP; the textual space for **T**, **CM** (text) and models with $g = id$; and the visual space for **CM** (vis.).

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

Classification benchmarks. All extrinsic evaluations are carried out using the SentEval pipeline (Conneau and Kiela, 2018). The tasks are the following: opinion polarity (MPQA) (Wiebe and Cardie, 2005), movie review sentiment (MR) (Pang and Lee, 2005), subjectivity/objectivity classification (SUBJ) (Pang and Lee, 2004), customer reviews (CR) (Hu and Liu, 2004), binary sentiment analysis on SST (Socher et al., 2013), paraphrase identification (MSRP) (Dolan et al., 2004) as well as two entailment classification benchmarks: SNLI (Bowman et al., 2015) and SICK (Marelli et al., 2014b). For each dataset, a logistic regression classifier is learned from the extracted sentence embeddings, and we report the classification accuracy.

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

$C_{inter} = \mathbb{E}_{v_s \neq v_{s'}}[cos(s, s')]$, which measures how well clusters are separated from each other.

4.4 Implementation details

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

5 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 (sect. 5.1), and quantitative results on the SentEval benchmark (sect. 5.2).

5.1 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 1, we report the structural measures and the semantic relatedness scores of the baselines, namely **T** and **CM**, and on the various scenarios of our model. The textual loss is discarded to isolate the effect of the different grounding hypotheses.

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

Grounded model	Textual model

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.

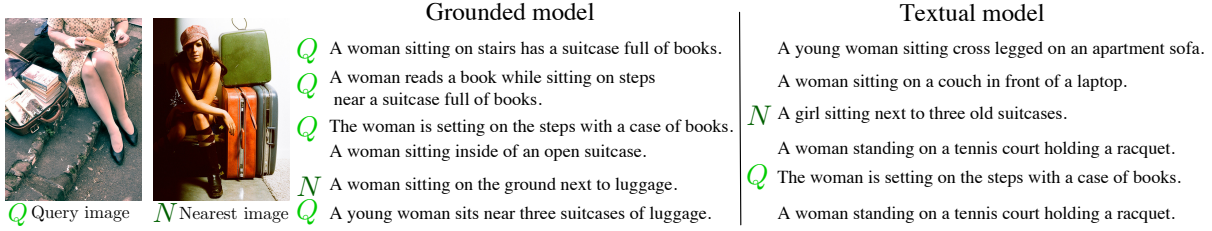Q Query image N Nearest image

Figure 2: Nearest neighbors of a selected sentence in the validation set of MS COCO, for both grounded and purely textual models. Q is the query image, N is the nearest neighbor of Q in the visual space. Sentences that are caption of Q or N are prefixed with Q or N.

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 2: Qualitative analysis: nearest neighbor of a given query (containing an abstract word) among Flickr30K sentences.
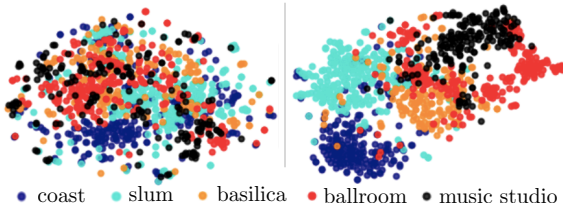

• coast • slum • basilica • ballroom • music studio

Figure 3: t-SNE visualization on CMPlaces sentences for a set of randomly sampled visual scenes. Left: textual model **T**. Right: grounded model $\mathbf{C}_g + \mathbf{P}_g$.

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

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

t-SNE visualization   This finding is also supported by a qualitative experiment showing 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 3 the corresponding sentences using t-SNE (Maaten and Hinton, 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 3: $C_{inter} = 19, C_{intra} = 22$ 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).

Nearest neighbors search   Furthermore, we show in Table 2 that concrete knowledge acquired via our grounded model can also be transferred to abstract sentences. To do so, we manually build sentences using words with low concreteness (between 2.5 and 3.5) from the USF dataset (Nelson et al., 2004). Then, nearest neighbors are retrieved from the set of sentences of Flickr30K (Plummer et al., 2015). In this sample, we see that our grounded model is more accurate than the purely textual model to capture visual meaning. The observation that visual information propagates from

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., 2015a)‡	**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 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. '†': the model has been re-implemented (we obtained higher scores than the one given in the original papers). '‡': the baseline is an adaptation of the model to the case of sentences. '∗': significantly differs from the best scenario among our models.

concrete sentences to abstract ones is analogous to findings made in previous research on word embeddings (Hill and Korhonen, 2014).

Neighboring structure   To illustrate the discrepancy on the mNNO metric observed between $\mathbf{C}_g + \mathbf{P}_g$ and $\mathbf{T}$, we select a query image Q in the validation set of MS COCO, along with its corresponding caption S; we display, in Figure 2, the nearest neighbor of Q in the visual space, noted N, and the nearest neighbors of S in the grounded space. With our grounded model, the neighborhood S is mostly made of sentences corresponding to Q or N.

Hypotheses validation   We now validate our hypotheses (cf. sect. 3.1) on the grounded space, using the Cross-Modal Projection baseline (**CM**) and our model scenarios as outlined in Table 1. For fair comparison, metrics for the baseline **CM** are estimated either on the visual or the textual space depending on whether our models rely on the grounded space (g) or not (id). These results correspond to the rows **CM** (text) and **CM** (vis.) in Table 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 1, e.g., $\mathbf{C}_g > \mathbf{C}_{id}$, $\mathbf{P}_g > \mathbf{P}_{id}$ and $\mathbf{C}_g + \mathbf{P}_g > \mathbf{C}_{id} + \mathbf{P}_{id}$; (2) Solely using cluster information leads to the highest C_{intra} and lowest C_{inter}, which suggests that $\mathbf{C}_\bullet$ is the most efficient model at separating visually different sentences; (3) Using only perceptual information in $\mathbf{P}_\bullet$ logically leads to highly correlated textual and visual spaces (highest ρ_{vis}), but the local neighborhood structure is not well preserved (lowest C_{intra}); (4) Our model $\mathbf{C}_\bullet + \mathbf{P}_\bullet$ is better than **CM** at capturing cluster information (higher C_{intra}, lower C_{inter}) and perceptual information (higher ρ_{vis}). This also translates in a higher mNNO measure for $\mathbf{C}_\bullet + \mathbf{P}_\bullet$, leading us to think that the conjunction of both perceptual and cluster information leads to high correlation of modalities, in terms of neighborhood structure. Moreover, this high mNNO score results in better performances for our model $\mathbf{C}_\bullet + \mathbf{P}_\bullet$ in terms of semantic relatedness.

5.2 Evaluation on transfer tasks

We now focus on extrinsic evaluation of the embeddings. Table 3 reports evaluations of our baselines and scenarios on SentEval (Conneau and Kiela, 2018), a classical benchmark used for evaluating sentence embeddings. Before further analysis, we find that our grounded models systematically outperform the textual baseline **T**, on all benchmarks, which shows the first substantial improvement brought by grounding and visual information in a sentence representation model. Indeed, models GS-Cap, GS-Img and GS-Both from (Kiela et al., 2018), despite improving over $\mathbf{T}_{1024}$, perform worse than the textual model of the same dimension **T** — this is consistent with what they report in their paper.

Our results interpretation is the following: (1) our joint approach shows superior performances over the sequential one, confirming results reported at the word level (Zablocki et al., 2018). 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 ($g = \text{MLP}$) can take advantage of more expression power provided by the trainable g than models which integrate grounded information directly in the textual space ($g = id$). (4) Among our model scenarios, $\mathbf{T} + \mathbf{P}_g$ has maximal scores on the most tasks; however, it shows lower scores on SNLI and SICK, which are entailment tasks. Models using cluster information $\mathbf{C}_g$ are naturally more suited for these tasks and hence obtain higher results. Finally, the combined model $\mathbf{T} + \mathbf{C}_g + \mathbf{P}_g$ shows a good balance between classification and entailment tasks.

6 Conclusion

We proposed a multimodal model aiming at preserving the structure of visual and textual spaces to learn grounded sentence representations. Our contributions include (1) leveraging both perceptual and cluster information and (2) using an intermediate grounded space enabling to relax the constraints on the textual space. Our approach is the first to report consistent positive results against purely textual baselines on a variety of natural language tasks. As future work, we plan to use visual information to specifically target complex downstream tasks requiring commonsense and reasoning such as question answering or visual dialogue.

Acknowledgments

References

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

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473.

Marco Baroni. 2016. Grounding distributional semantics in the visual world. *Language and Linguistics Compass*, 10(1):3–13.

Samuel R. Bowman, 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*, pages 632–642.

Elia Bruni, Nam Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *J. Artif. Int. Res.*, 49(1):1–47.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2013. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46.

Curt Burgess and Kevin Lund. 1997. Modelling parsing constraints with high-dimensional context space. 12.

Micael Carvalho, 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*, pages 35–44.

Lluis Castrejon, Yusuf Aytar, Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. 2016. Learning aligned cross-modal representations from weakly aligned data. In *Computer Vision and Pattern Recognition (CVPR), 2016 IEEE Conference on*. IEEE.

Daniel Cer, 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*, pages 169–174.

Daniel M. Cer, 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*, pages 1–14.

Noam Chomsky. 1980. Rules and representations. *Behavioral and brain sciences*, 3(1):1–15.

Grzegorz Chrupala, 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*, pages 112–118.

Guillem Collell and Marie-Francine Moens. 2018. 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*, pages 462–468.

Guillem Collell, Teddy Zhang, and Marie-Francine Moens. 2017. Imagined visual representations as multimodal embeddings. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*. AAAI.

Alexis Conneau 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*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc 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*, pages 670–680.

Bill Dolan, 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*.

Rebecca Fincher-Kiefer. 2001. Perceptual components of situation models. *Memory & Cognition*, 29(2):336–343.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction*, AKBC '13, pages 25–30, New York, NY, USA. ACM.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

Felix Hill, 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*, pages 1367–1377.

Felix Hill and Anna Korhonen. 2014. 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*, pages 255–265.

Felix Hill, Roi Reichart, and Anna Korhonen. 2014. Multi-modal models for concrete and abstract concept meaning. *Transactions of the Association for Computational Linguistics*, 2:285–296.

Minqing Hu 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, pages 168–177, New York, NY, USA. ACM.

Nal Kalchbrenner, 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*, pages 655–665.

Andrej Karpathy 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*, pages 3128–3137.

Douwe Kiela, 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)*, pages 408–418.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Ryan Kiros, 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*, pages 3294–3302.

Satwik Kottur, 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*, pages 4985–4994.

Angeliki Lazaridou, Georgiana Dinu, and Marco Baroni. 2015a. Hubness and pollution: Delving into cross-space mapping for zero-shot learning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 270–280. Association for Computational Linguistics.

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015b. 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*, pages 153–163.

Tsung-Yi Lin, 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*, pages 740–755.

Xiao Lin 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*, pages 2984–2993.

Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *CoRR*, abs/1703.03130.

Lajanugen Logeswaran 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*.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

Marco Marelli, 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.*, pages 1–8.

Marco Marelli, 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.*, pages 216–223.

Tomas Mikolov, 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.*, pages 3111–3119.

Douglas L Nelson, Cathy L McEvoy, and Thomas A Schreiber. 2004. The university of south florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, 36(3):402–407.

Donald A Norman. 1972. Memory, knowledge, and the answering of questions.

Bo Pang 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.*, pages 271–278.

Bo Pang 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*, pages 115–124.

Matthew E. Peters, 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)*, pages 2227–2237.

Bryan A. Plummer, 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*, pages 2641–2649.

Carina Silberer 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*, pages 721–732.

Richard Socher, 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, pages 1631–1642.

Christian Szegedy, 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*, pages 2818–2826.

Lawrence W. Barsalou. 1999. Perceptual symbol systems. 22:577–609; discussion 610.

Hong Wang, Wenhan Xiong, Mo Yu, Xiaoxiao Guo, Shiyu Chang, and William Yang Wang. 2019. Sentence embedding alignment for lifelong relation extraction. *NAACL*.

Liwei Wang, Yin Li, and Svetlana Lazebnik. 2016. Learning deep structure-preserving image-text embeddings. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 5005–5013.

Janyce Wiebe 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*, page 2005.

Fanyi Xiao, Leonid Sigal, and Yong Jae Lee. 2017. Weakly-supervised visual grounding of phrases with linguistic structures. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5253–5262.

Mark Yatskar, 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*, pages 193–198.

Éloi Zablocki, Benjamin Piwowarski, Laure Soulier, and Patrick Gallinari. 2018. Learning multi-modal word representation grounded in visual context. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018*.

707