Learning Visually Grounded Sentence Representations

Douwe Kiela
Facebook AI Research
dkiela@fb.com

Alexis Conneau
Facebook AI Research
aconneau@fb.com

Allan Jabri[1]
UC Berkeley
ajabri@berkeley.edu

Maximilian Nickel
Facebook AI Research
maxn@fb.com

Abstract

We investigate grounded sentence representations, where we train a sentence encoder to predict the image features of a given caption—i.e., we try to "imagine" how a sentence would be depicted visually—and use the resultant features as sentence representations. We examine the quality of the learned representations on a variety of standard sentence representation quality benchmarks, showing improved performance for grounded models over non-grounded ones. In addition, we thoroughly analyze the extent to which grounding contributes to improved performance, and show that the system also learns improved word embeddings.

1 Introduction

Following the word embedding upheaval of the past few years, one of NLP's next big challenges has become the hunt for universal sentence representations: generic representations of sentence meaning that can be "plugged into" any kind of system or pipeline. Examples include Paragraph2Vec (Le and Mikolov, 2014), C-Phrase (Pham et al., 2015), SkipThought (Kiros et al., 2015) and FastSent (Hill et al., 2016a). These representations tend to be learned from large corpora in an unsupervised setting, much like word embeddings, and effectively "transferred" to the task at hand.

Purely text-based semantic models, which represent word meaning as a distribution over other words (Harris, 1954; Turney and Pantel, 2010; Clark, 2015), suffer from the grounding problem (Harnad, 1990). It has been shown that grounding leads to improved performance on a variety of word-level tasks (Baroni, 2016; Kiela, 2017). Unsupervised sentence representation models are often doubly exposed to the grounding problem, especially if they represent sentence meanings as a distribution over other sentences, as in SkipThought (Kiros et al., 2015).

Here, we examine whether grounding also leads to improved sentence representations. In short, the grounding problem is characterized by the lack of an association between symbols and external information. We address this problem by aligning text with paired visual data and hypothesize that sentence representations can be enriched with external information—i.e., grounded—by forcing them to capture visual semantics. We investigate the performance of these representations and the effect of grounding on a variety of semantic benchmarks.

There has been much recent interest in generating actual images from text (Goodfellow et al., 2014; van den Oord et al., 2016; Mansimov et al., 2016). Our method takes a slightly different approach: instead of predicting actual images, we train a deep recurrent neural network to predict the *latent feature representation* of images. That is, we are specifically interested in the semantic content of visual representations and how useful that information is for learning sentence representations. One can think of this as trying to imagine, or form a "mental picture", of a sentence's meaning (Chrupała et al., 2015). Much like a sentence's meaning in classical semantics is given by its model-theoretic ground truth (Tarski, 1944), our ground truth is provided by images.

Grounding is likely to be more useful for concrete words and sentences: a sentence such as "democracy is a political system" does not yield any coherent mental picture. In order to accommodate the fact that much of language is abstract, we take sentence representations obtained using text-only data (which are better for representing abstract meaning) and combine them with the grounded representations that our system learns (which are good for representing concrete meaning), leading to multi-modal sentence representations.

[1]Work done while at Facebook AI Research.

Proceedings of NAACL-HLT 2018, pages 408–418
New Orleans, Louisiana, June 1 - 6, 2018.

In what follows, we introduce a system for grounding sentence representations by learning to predict visual content. Although it is not the primary aim of this work, it is important to first examine how well this system achieves what it is trained to do, by evaluating on the COCO5K image and caption retrieval task. We then analyze the performance of grounded representations on a variety of sentence-level semantic transfer tasks, showing that grounding increases performance over text-only representations. We then investigate an important open question in multi-modal semantics: to what extent are improvements in semantic performance due to grounding, rather than to having more data or data from a different distribution? In the remainder, we analyze the role that concreteness plays in representation quality and show that our system learns grounded word embedding projections that outperform non-grounded ones. To the best of our knowledge, this is the first work to comprehensively study grounding for distributed sentence representations on such a wide set of semantic benchmark tasks.

2 Related work

Sentence representations Although there appears to be a consensus with regard to the methodology for learning word representations, this is much more of an open problem for sentence representations. Recent work has ranged from trying to learn to compose word embeddings (Le and Mikolov, 2014; Pham et al., 2015; Wieting et al., 2016; Arora et al., 2017), to neural architectures for predicting the previous and next sentences (Kiros et al., 2015) or learning representations via large-scale supervised tasks (Conneau et al., 2017). In particular, SkipThought (Kiros et al., 2015) led to an increased interest in learning sentence representations. Hill et al. (2016a) compare a wide selection of unsupervised and supervised methods, including a basic caption prediction system that is similar to ours. That study finds that "different learning methods are preferable for different intended applications", i.e., that the matter of optimal universal sentence representations is as of yet far from decided.

InferSent (Conneau et al., 2017) recently showed that supervised sentence representations can be of very high quality. Here, we learn grounded sentence representations in a supervised setting, combine them with standard unsupervised sentence representations, and show how grounding can help for a variety of sentence-level tasks.

Multi-modal semantics Language grounding in semantics has been motivated by evidence that human meaning representations are grounded in perceptual experience (Jones et al., 1991; Perfetti, 1998; Andrews et al., 2009; Riordan and Jones, 2011). That is, despite ample evidence of humans representing meaning with respect to an external environment and sensorimotor experience (Barsalou, 2008; Louwerse, 2008), standard semantic models rely solely on textual data. This gives rise to an infinite regress in text-only semantic representations, i.e., words are defined in terms of other words, *ad infinitum*.

The field of multi-modal semantics, which aims to address this issue by enriching textual representations with information from other modalities, has mostly been concerned with word representations (Bruni et al., 2014; Baroni, 2016; Kiela, 2017, and references therein). Learning multi-modal representations that ground text-only representations has been shown to improve performance on a variety of core NLP tasks. This work is most closely related to that of Chrupała et al. (2015), who also aim to ground language by relating images to captions: here, we additionally address abstract sentence meaning; have a different architecture, loss function and fusion strategy; and explicitly focus on grounded universal sentence representations.

Bridging vision and language There is a large body of work that involves jointly embedding images and text, at the word level (Frome et al., 2013; Joulin et al., 2016), the phrase level (Karpathy et al., 2014; Li et al., 2016), and the sentence level (Karpathy and Fei-Fei, 2015; Klein et al., 2015; Kiros et al., 2015; Chen and Zitnick, 2015; Reed et al., 2016). Our model similarly learns to map sentence representations to be consistent with a visual semantic space, and we focus on studying how these grounded text representations transfer to NLP tasks.

Moreover, there has been a lot of work in recent years on the task of image caption generation (Bernardi et al., 2016; Vinyals et al., 2015; Mao et al., 2015; Fang et al., 2015). Here, we do the opposite: we predict the correct image (features) from the caption, rather than the caption from the image (features). Similar ideas were recently successfully applied to multi-modal machine translation

(Elliott and Kádár, 2017; Gella et al., 2017; Lee et al., 2017). Recently, Das et al. (2017) trained dialogue agents to communicate about images, trying to predict image features as well.

3 Approach

In the following, let $\mathcal{D} = \{(I_k, C_k)\}_{k=1}^{N}$ be a dataset where each image I_k is associated with one or more captions $C_k = \{C_1, \ldots, C_{|C|_k}\}$. A prominent example of such a dataset is COCO (Lin et al., 2014), which consists of images with up to 5 corresponding captions for each image. The objective of our approach is to encode a given sentence, i.e., a caption C, and learn to ground it in the corresponding image I. To encode the sentence, we train a bidirectional LSTM (BiLSTM) on the caption, where the input is a sequence of projected word embeddings. We combine the final left-to-right and right-to-left hidden states of the LSTM and take the element-wise maximum to obtain a sentence encoding. We then examine three distinct methods for grounding the sentence encoding.

In the first method, we try to predict the image features (Cap2Img). That is, we learn to map the caption to the same space as the image features that represent the correct image. We call this strong perceptual grounding, where we take the visual input directly into account.

An alternative method is to exploit the fact that one image in COCO has multiple captions (Cap2Cap), and to learn to predict which other captions are valid descriptions of the same image. This approach is strictly speaking not perceptually grounded, but exploits the fact that there is an implicit association between the captions and the shared underlying image, and so could be considered a weaker version of grounding.

Finally, we experiment with a model that optimizes both these objectives jointly: that is, we predict both images and alternative captions for the same image (Cap2Both). Thus, Cap2Both incorporates both strong perceptual and weak implicit grounding. Please see Figure 1 for an illustration of the various models. In what follows, we discuss them in more technical detail.

3.1 Bidirectional LSTM

To learn sentence representations, we employ a bidirectional LSTM architecture. In particular, let $x = (x_1, \ldots, x_T)$ be an input sequence where each word is represented via an embedding $\mathbf{x}_t \in \mathbb{R}^n$. Using a standard LSTM (Hochreiter and Schmidhuber, 1997), the hidden state at time t, denoted $\mathbf{h}_t \in \mathbb{R}^m$, is computed via

$$\mathbf{h}_{t+1}, \mathbf{c}_{t+1} = \text{LSTM}(\mathbf{x}_t, \mathbf{h}_t, \mathbf{c}_t \mid \Theta)$$

where $\mathbf{c}_t$ denotes the cell state of the LSTM and where Θ denotes its parameters.

To exploit contextual information in both input directions, we process input sentences using a bidirectional LSTM, that reads an input sequence in both normal and reverse order. In particular, for an input sequence x of length T, we compute the hidden state at time t, $\mathbf{h}_t \in \mathbb{R}^{2m}$ via

$$\mathbf{h}^f_{t+1} = \text{LSTM}(\mathbf{x}_t, \mathbf{h}^f_t, \mathbf{c}^f_t \mid \Theta^f)$$
$$\mathbf{h}^b_{t+1} = \text{LSTM}(\mathbf{x}_{T-t}, \mathbf{h}^b_t, \mathbf{c}^b_t \mid \Theta^b)$$

Here, the two LSTMs process x in a forward and a backward order, respectively. We subsequently use $\max : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^d$ to combine them into their element-wise maximum, yielding the representation of a caption after it has been processed with the BiLSTM:

$$\mathbf{h}_T = \max(\mathbf{h}^f_t, \mathbf{h}^b_t)$$

We use GloVe vectors (Pennington et al., 2014) for our word embeddings. The embeddings are kept fixed during training, which allows a trained sentence encoder to transfer to tasks (and a vocabulary) that it has not yet seen, provided GloVe embeddings are available. Since GloVe representations are not tuned to represent grounded information, we learn a global transformation of GloVe space to grounded word space. Specifically, let $\overline{\mathbf{x}} \in \mathbb{R}^n$ be the original GloVe embeddings. We then learn a linear map $U \in \mathbb{R}^{n \times n}$ such that $\mathbf{x} = U\overline{\mathbf{x}}$ and use $\mathbf{x}$ as input to the BiLSTM. The linear map U and the BiLSTM are trained jointly.

3.2 Cap2Img

Let $\mathbf{v} \in \mathbb{R}^I$ be the latent representation of an image (e.g.the final layer of a ResNet). To ground captions in the images that they describe, we map $\mathbf{h}_T$ into the latent space of image representations such that their similarity is maximized. In other words, we aim to predict the latent features of an image from its caption. The mapping of caption to image space is performed via a series of projections

$$\mathbf{p}_0 = \mathbf{h}_T$$
$$\mathbf{p}_{\ell+1} = \psi(P_\ell \mathbf{p}_\ell)$$

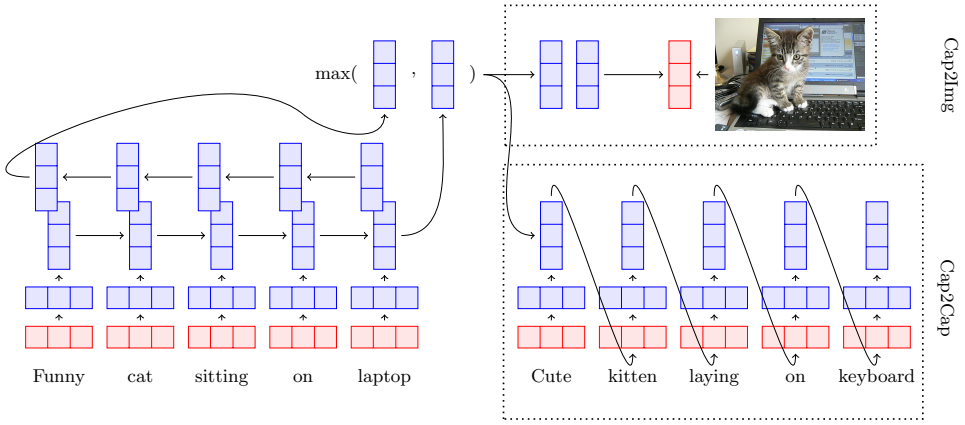

Figure 1: Model architecture: predicting either an image (Cap2Img), an alternative caption (Cap2Cap), or both at the same time (Cap2Both).

where ψ denotes a non-linearity such as ReLUs or tanh.

By jointly training the BiLSTM with these latent projections, we can then ground the language model in its visual counterpart. In particular, let $\Theta = \Theta_{\text{BiLSTM}} \cup \{P_\ell\}_{\ell=1}^{L}$ be the parameters of the BiLSTM as well as the projection layers. We then minimize the following ranking loss:

$$\mathcal{L}_{\text{C2I}}(\Theta) = \sum_{(I,C)\in\mathcal{D}} f_{\text{rank}}(I, C) + f_{\text{rank}}(C, I) \quad (1)$$

where

$$f_{\text{rank}}(a, b) = \sum_{b'\in\mathcal{N}_a} [\gamma - \text{sim}(a, b) + \text{sim}(a, b')]_+$$

where $[x]_+ = \max(0, x)$ denotes the threshold function at zero and γ defines the margin. Furthermore, $\mathcal{N}_a$ denotes the set of negative samples for an image or caption and $\text{sim}(\cdot, \cdot)$ denotes a similarity measure between vectors. In the following, we employ the cosine similarity, i.e.,

$$\text{sim}(a, b) = \frac{\langle \mathbf{a}, \mathbf{b} \rangle}{\|\mathbf{a}\|\|\mathbf{b}\|}.$$

Although this loss is not smooth at zero, it can be trained end-to-end using subgradient methods. Compared to e.g. an l_2 regression loss, Equation (1) is less susceptible to error incurred by subspaces of the visual representation that are irrelevant to the high level visual semantics. Empirically, we found it to be more robust to overfitting.

3.3 Cap2Cap

Let $x = (x_1, \ldots, x_T)$, $y = (y_1, \ldots, y_S)$ be a caption pair that describes the same image. To learn weakly grounded representations, we employ a standard sequence-to-sequence model (Sutskever et al., 2014), whose task is to predict y from x. As in the Cap2Cap model, let $\mathbf{h}_T$ be the representation of the input sentence after it has been processed with a BiLSTM. We then model the joint probability of y given x as

$$p(y \mid x) = \prod_{s=1}^{S} p(y_s \mid \mathbf{h}_T, y_1, \ldots, y_{s-1}, \Theta).$$

To model the conditional probability of y_s we use the usual multiclass classification approach over the vocabulary of the corpus $\mathcal{V}$ such that

$$p(y_s = k \mid \mathbf{h}_T, y_1, \ldots, y_{s-1}, \Theta) = \frac{e^{\langle \mathbf{v}_k, \mathbf{y}_s \rangle}}{\sum_{j=1}^{|\mathcal{V}|} e^{\langle \mathbf{v}_j, \mathbf{y}_s \rangle}}.$$

Here, $\mathbf{y}_s = \psi(W_V \mathbf{g}_s + \mathbf{b})$ and $\mathbf{g}_s$ is hidden state of the decoder LSTM at time s.

To learn the model parameters, we minimize the negative log-likelihood over all caption pairs, i.e.,

$$\mathcal{L}_{\text{C2C}}(\theta) = -\sum_{x,y\in\mathcal{D}} \sum_{s=1}^{|y|} \log p(y_s \mid \mathbf{h}_T, y_1, \ldots, y_{s-1}, \Theta).$$

3.4 Cap2Both

Finally, we also integrate both concepts of grounding into a joint model, where we optimize the following loss function:

$$\mathcal{L}_{\text{C2B}}(\Theta) = \mathcal{L}_{C2I}(\Theta) + \mathcal{L}_{C2C}(\Theta).$$

3.5 Grounded universal representations

On their own, features from this system are likely to suffer from the fact that training on COCO introduces biases: aside from the inherent dataset bias in COCO itself, the system will only have coverage for concrete concepts. COCO is also a much smaller dataset than e.g. the Toronto Books Corpus often used in purely text-based methods (Kiros et al., 2015). As such, grounded representations are potentially less "universal" than text-based alternatives, which also cover abstract concepts.

There is evidence that meaning is dually coded in the human brain: while abstract concepts are processed in linguistic areas, concrete concepts are processed in both linguistic and visual areas (Paivio, 1990). Anderson et al. (2017) recently corroborated this hypothesis using semantic representations and fMRI studies. In our case, we want to be able to accommodate concrete sentence meanings, for which our vision-centric system is likely to help; as well as abstract sentence meanings, where trying to "imagine" what "democracy is a political system" might look like will probably only introduce noise.

Hence, we optionally complement our systems' representations with more abstract universal sentence representations trained on language-only data (specifically, the Toronto Books Corpus). Although it would be interesting to examine multitask scenarios where these representations are jointly learned, we leave this for future work. Here, instead, we combine grounded and language-only representations using simple concatenation, i.e., $r_{gs} = r_{grounded} \mid\mid r_{ling-only}$. Concatenation has been proven to be a strong and straightforward mid-level multi-modal fusion method, previously explored in multi-modal semantics for word representations (Bruni et al., 2014; Kiela and Bottou, 2014). We call the combined system GroundSent (GS), and distinguish between sentences perceptually grounded in images (GroundSent-Img), weakly grounded in captions (GroundSent-Cap) or grounded in both (GroundSent-Both).

3.6 Implementation details

We use 300-dimensional GloVe (Pennington et al., 2014) embeddings, trained on WebCrawl, for the initial word representations and optimize using Adam (Kingma and Ba, 2015). We use ELU (Clevert et al., 2016) for the non-linearity in projection layers, set dropout to 0.5 and use a dimensionality of 1024 for the LSTM. The network was initialized with orthogonal matrices for the recurrent layers (Saxe et al., 2014) and He initialization (He et al., 2015) for all other layers. The learning rate and margin were tuned on the validation set using grid search.

4 Data, evaluation and comparison

We use the same COCO splits as Karpathy and Fei-Fei (2015) for training (113,287 images), validation (5000 images) and testing (5000 images). Image features for COCO were obtained by transferring the final layer from a ResNet-101 (He et al., 2016) trained on ImageNet (ILSVRC 2015).

4.1 Transfer tasks

We are specifically interested in how well (grounded) universal sentence representations transfer to different tasks. To evaluate this, we perform experiments for a variety of tasks. In all cases, we compare against layer-normalized Skip-Thought vectors, a well-known high-performing sentence encoding method (Ba et al., 2016). To ensure that we use the exact same evaluations, with identical hyperparameters and settings, we evaluate all systems with the same evaluation pipeline, namely SentEval (Conneau and Kiela, 2018)[2]. Following previous work in the field, the idea is to take universal sentence representations and to learn a simple classifier on top for each of the transfer tasks—the higher the quality of the sentence representation, the better the performance on these transfer tasks should be.

4.1.1 Semantic classification

We evaluate on the following well-known and widely used evaluations: movie review sentiment (MR) (Pang and Lee, 2005), product reviews (CR) (Hu and Liu, 2004), subjectivity classification (SUBJ) (Pang and Lee, 2004), opinion polarity (MPQA) (Wiebe et al., 2005), paraphrase identification (MSRP) (Dolan et al., 2004) and sentiment classification (SST, binary version) (Socher et al., 2013). Accuracy is measured in all cases, except for MRPC, which measures accuracy and the F1-score.

[2]See https://github.com/facebookresearch/SentEval. The aim of SentEval is to encompass a comprehensive set of benchmarks that has been loosely established in the research community as the standard for evaluating sentence representations.

	COCO5K									
	Caption Retrieval					Image Retrieval				
Model	**R@1**	**R@5**	**R@10**	**MEDR**	**MR**	**R@1**	**R@5**	**R@10**	**MEDR**	**MR**
DVSA	11.8	32.5	45.4	12.2	NA	8.9	24.9	36.3	19.5	NA
FV	17.3	39.0	50.2	10.0	46.4	10.8	28.3	40.1	17.0	49.3
OE	23.3	NA	65.0	5.0	24.4	18.0	NA	57.6	7.0	35.9
Cap2Both	19.4	45.0	59.4	7.0	26.5	11.7	32.6	46.4	12.0	41.7
Cap2Img	27.1	55.6	70.0	4.0	19.2	17.1	43.0	57.3	8.0	36.6

Table 1: Retrieval (higher is better) results on COCO, plus median rank (MEDR) and mean rank (MR) (lower is better). Note that while this work underwent review, better methods have been published, most notably VSE++ (Faghri et al., 2017).

Model	**MR**	**CR**	**SUBJ**	**MPQA**	**MRPC**	**SST**	**SNLI**	**SICK**
ST-LN	78.1	80.1	92.7	88.0	69.6/81.2	82.9	73.8	78.5
GroundSent-Cap	79.9	81.4	93.1	88.9	72.9/82.2	85.0	75.5	79.7
GroundSent-Img	79.1	80.8	93.1	89.0	71.9/81.4	86.1	76.1	82.2
GroundSent-Both	79.6	81.7	93.4	89.4	72.7/82.5	84.8	76.1	81.6

Table 2: Accuracy results on sentence classification and entailment tasks.

4.1.2 Entailment

Recent years have seen an increased interest in entailment classification as an appropriate evaluation of sentence representation quality. We evaluate representations on two well-known entailment, or natural language inference, datasets: the large-scale SNLI dataset (Bowman et al., 2015) and the SICK dataset (Marelli et al., 2014).

4.2 Implementational details

We implement a simple logistic regression on top of the sentence representation. In the cases of SNLI and SICK, as is the standard for these datasets, the representations for the individual sentences u and v are combined by using $\langle \mathbf{u}, \mathbf{v}, \mathbf{u} * \mathbf{v}, |\mathbf{u} - \mathbf{v}| \rangle$ as the input features. We tune the seed and an l_2 penalty on the validation sets for each, and train using Adam (Kingma and Ba, 2015), with a learning rate of 0.001 and a batch size of 32.

5 Results

Although it is not the primary aim of this work to learn a state-of-the-art image and caption retrieval system, it is important to first establish the capability of our system to do what it is trained to do. Table 1 shows the results on the COCO5K caption and image retrieval tasks for the two models that predict image features. We compare our system against several well-known approaches, namely Deep Visual-Semantic Alignments (DVSA) (Karpathy and Fei-Fei, 2015), Fisher Vectors (FV) (Klein et al., 2015) and Order Embeddings (OE) (Vendrov et al., 2015). As the results show, Cap2Img performs very well on this task, outperforming the compared models on caption retrieval and being very close to order embeddings on image retrieval[3]. The fact that the system outperforms Order Embeddings on caption retrieval suggests that it has a better sentence encoder. Cap2Both does not work as well on this task as the image-only case, probably because interference from the language signal makes the problem harder to optimize. The results indicate that the system has learned to predict image features from captions, and captions from images, at a level exceeding or close to the state-of-the-art on this task.

5.1 Transfer task performance

Having established that we can learn high-quality grounded sentence encodings, the core question we now wish to examine is how well grounded sentence representations transfer. In this section, we combine our grounded features with the

[3]In fact, we found that we can achieve better performance on this task by reducing the dimensionality of the encoder. A lower dimensionality in the encoder also reduces the transferability of the features, unfortunately, so we leave a more thorough investigation of this phenomenon for future work.

Model	MR	CR	SUBJ	MPQA	MRPC	SST	SNLI	SICK
STb-1024	70.3	68.0	87.5	85.5	69.7/80.6	78.3	67.3	76.6
STb-2048	73.1	**75.7**	88.3	86.5	71.6/**81.7**	79.0	71.0	78.8
2×STb-1024	71.4	74.7	88.2	86.6	71.3/80.7	75.8	69.4	78.3
Cap2Cap	71.4	74.7	86.7	**86.7**	70.3/79.8	76.1	68.5	78.2
Cap2Img	72.1	75.5	86.9	86.0	**72.3**/81.1	77.7	71.4	81.2
Cap2Both	71.6	74.4	86.5	85.5	71.4/79.5	78.5	71.3	**81.7**
GroundSent-Cap	73.1	73.0	**88.6**	86.6	70.8/81.2	79.4	70.7	79.1
GroundSent-Img	72.5	74.9	88.4	85.7	71.3/81.2	79.4	70.5	79.7
GroundSent-Both	**73.3**	75.2	87.5	86.6	69.9/79.9	**80.3**	**72.0**	78.1

Table 3: Thorough investigation of the contribution of grounding, ensuring equal number of components and identical architectures, on the variety of sentence-level semantic benchmark tasks. STb=SkipThought-like model with bidirectional LSTM+max. 2×STb-1024=ensemble of 2 different STb models with different initializations. GroundSent is STb-1024+Cap2Cap/Img/Both. We find that performance improvements are sometimes due to having more parameters, but in most cases due to grounding.

high-quality layer-normalized SkipThought representations of Ba et al. (2016), leading to multi-modal sentence representations as described in Section 3.5. That is, we concatenate Cap2Cap, Cap2Img or Cap2Both and Skip-Thought with Layer Normalization (ST-LN) representations, yielding GroundSent-Cap, GroundSent-Img and GroundSent-Both representations, respectively. We report performance of ST-LN using SentEval, which led to slightly different numbers than what is reported in their paper[4].

Table 2 shows the results for the semantic classification and entailment tasks. Note that all systems use the exact same evaluation pipeline, which makes them directly comparable. We can see that in all cases, grounding increases the performance. The question of which type of grounding works best is more difficult: generally, grounding with Cap2Cap and Cap2Both appears to do slightly better on most tasks, but on e.g. SST, Cap2Img works better. The entailment task results (SNLI and SICK in Table 2) show a similar picture: in all cases grounding improves performance.

It is important to note that, in this work, we are not necessarily concerned with replacing the state-of-the-art on these tasks: there are systems that perform better. We are primarily interested in whether grounding helps relative to text-only baselines. We find that it does.

[4]This is probably due to different seeds, optimization methods and other minor implementational details that differ between the original work and SentEval.

5.2 The contribution of grounding

An important open question is whether the increase in performance in multi-modal semantic models is due to qualitatively different information from *grounding*, or simply due to the fact that we have *more parameters* or *data from a different distribution*. In order to examine this, we implement a SkipThought-like model that also uses a bidirectional LSTM with element-wise max on the final hidden layer (henceforth referred to as STb). This model is architecturally identical to the sentence encoder used before: it can be thought of as Cap2Cap, but where the objective is not to predict an alternative caption, but to predict the previous and next sentence in the Toronto Books Corpus, just like SkipThought (Kiros et al., 2015).

We train a 1024-dimensional and 2048-dimensional STb model (for one full iteration, with all other hyperparameters identical to Cap2Cap) to compare against: if grounding improves results because it introduces qualitatively different information, rather than just from having more parameters (i.e., a higher embedding dimensionality), we should expect the multi-modal GroundSent models to perform better not only than STb-1024, but also than STb-2048, which has the same number of parameters (recall that GroundSent models are combinations of grounded and linguistic-only representations). In addition, we compare against an "ensemble" of two different STb-1024 models (i.e., a concatenation of two separately trained STb-1024), to check that we are not (just) observing an ensemble effect.

Dataset	Concreteness
MR	2.3737 ± 0.965
CR	2.4714 ± 1.025
SUBJ	2.4510 ± 1.007
MPQA	2.3158 ± 0.834
MRPC	2.5086 ± 0.987
SST	2.7471 ± 1.142
SNLI	3.1867 ± 1.309
SICK	3.1282 ± 1.372

Table 4: Mean and variance of dataset concreteness, over all words in the datasets.

As Table 3 shows, a more nuanced picture emerges in this comparison: grounding helps more for some datasets than for others. Grounded models outperform the STb-1024 model (which uses much more data—the Toronto Books Corpus is much larger than COCO) in all cases, often already without concatenating the textual modality. The ensemble of two STb-1024 models performs better than the individual one, and so does the higher-dimensional one. In the cases of CR and MRPC (F1), it appears that improved performance is due to having more data or ensemble effects. For the other datasets, grounding clearly yields better results. These results indicate that grounding does indeed capture qualitatively different information, yielding better universal sentence representations.

6 Discussion

There are a few other important questions to investigate. The average abstractness or concreteness of the evaluation datasets may have a large impact on performance. In addition, word embeddings from the learned projection from GloVe input embeddings, which now provides a generic word-embedding grounding method even for words that are not present in the image-caption training data, can be examined.

6.1 Concreteness

As we have seen, performance across datasets and models can vary substantially. A dataset's concreteness plays an important role in the relative merit of applying grounding: a dataset consisting mostly of abstract words is less likely to benefit from grounding than one that uses mostly concrete words. In order to examine this effect, we calculate the average concreteness of the evaluation datasets used in this study. Table 4 shows the average human-annotated concreteness ratings for all words (where available) in each dataset. The ratings were obtained by Brysbaert et al. (2014) in a large-scale study, yielding scores for 40,000 English words.

We observe that the two entailment datasets are more concrete, which is due to the fact that the premises are derived from caption datasets (Flickr30K in the case of SNLI; Flickr8K and video captions in the case of SICK). This explains why grounding can clearly be seen to help in these cases. For the semantic classification tasks, the more concrete datasets are MRPC and SST. The picture is less clear for the first, but in SST we see that the grounded representations definitely do work better. Concreteness values make it easier to analyze performance, but are apparently not always direct indicators of improvements with grounding.

Model	MEN	SimLex	RW	W353
GloVe	0.805	0.408	0.451	0.738
Cap2Both	0.819	0.467	0.487	0.712
Cap2Img	0.845	0.515	0.523	0.753

Table 5: Spearman ρ_s correlation on four standard semantic similarity evaluation benchmarks.

6.2 Grounded word embeddings

Our models contain a projection layer that maps the GloVe word embeddings that they receive as inputs to a different embedding space. There has been a lot of interest in grounded word representations in recent years, so it is interesting to examine what kind of word representations our models learn. We omit Cap2Cap for reasons of space (it performs similarly to Cap2Both). As shown in Table 5, the grounded word projections that our network learns yield higher-quality word embeddings on four standard lexical semantic similarity benchmarks: MEN (Bruni et al., 2014), SimLex-999 (Hill et al., 2016b), Rare Words (Luong et al., 2013) and WordSim-353 (Finkelstein et al., 2001).

7 Conclusion

We have investigated grounding for universal sentence representations. We achieved good performance on caption and image retrieval tasks on the large-scale COCO dataset. We subsequently showed how the sentence encodings that the sys-

tem learns can be transferred to various NLP tasks, and that grounded universal sentence representations lead to improved performance. We analyzed the source of improvements from grounding, and showed that the increased performance appears to be due to the introduction of qualitatively different information (i.e., grounding), rather than simply having more parameters or applying ensemble methods. Lastly, we showed that our systems learned high-quality grounded word embeddings that outperform non-grounded ones on standard semantic similarity benchmarks. It could well be that our methods are even more suited for more concrete tasks, such as visual question answering, visual storytelling, or image-grounded dialogue—an avenue worth exploring in future work. In addition, it would be interesting to explore multi-task learning for sentence representations where one of the tasks involves grounding.

Acknowledgments

We thank the anonymous reviewers for their helpful comments and suggestions. Part of Fig. 1 is licensed from dougwoods/CC-BY-2.0/flickr.com/photos/deerwooduk/682390157.

References

Andrew J. Anderson, Douwe Kiela, Stephen Clark, and Massimo Poesio. 2017. Visually grounded and textual semantic models differentially decode brain activity associated with concrete and abstract nouns. *Transactions of the Association for Computational Linguistics* 5.

Mark Andrews, Gabriella Vigliocco, and David Vinson. 2009. Integrating experiential and distributional data to learn semantic representations. *Psychological review* 116(3):463.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. In *International Conference on Learning Representations (ICLR)*.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450* .

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

Lawrence W. Barsalou. 2008. Grounded cognition. *Annual Review of Psychology* 59(1):617–645.

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research (JAIR)* pages 409–442.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of EMNLP*.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artifical Intelligence Research* 49:1–47.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods* 46(3):904–911.

Xinlei Chen and Lawrence C Zitnick. 2015. Mind's eye: A recurrent visual representation for image caption generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. pages 2422–2431.

Grzegorz Chrupała, Ákos Kádár, and Afra Alishahi. 2015. Learning language through pictures. In *Proceedings of ACL*.

Stephen Clark. 2015. Vector Space Models of Lexical Meaning. In Shalom Lappin and Chris Fox, editors, *Handbook of Contemporary Semantic Theory*, Wiley-Blackwell, Oxford, chapter 16.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. 2016. Fast and accurate deep network learning by exponential linear units (ELUs). In *International Conference on Learning Representations (ICLR)*.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. In *Proceedings of LREC*.

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 EMNLP*. Copenhagen, Denmark.

Abhishek Das, Satwik Kottur, José M. F. Moura, Stefan Lee, and Dhruv Batra. 2017. Learning cooperative visual dialog agents with deep reinforcement learning. In *Proceedings of CVPR*.

Bill Dolan, Chris Quirk, and Chris Brockett. 2004. Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceedings of ACL*. page 350.

Desmond Elliott and Ákos Kádár. 2017. Imagination improves multimodal translation. *arXiv preprint arXiv:1705.04350* .

Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2017. Vse++: Improved visual-semantic embeddings. *arXiv preprint arXiv:1707.05612* .

H. Fang, S. Gupta, F.N. Iandola, R. Srivastava, L. Deng, P. Dollar, J. Gao, X. He, M. Mitchell, J.C. Platt, C.L. Zitnick, and G. Zweig. 2015. From captions to visual concepts and back. In *CVPR*.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. In *Proceedings of the 10th international conference on World Wide Web (WWW)*. ACM, pages 406–414.

A. Frome, G. Corrado, J. Shlens, S. Bengio, J. Dean, and T. Mikolov. 2013. Devise: A deep visual-semantic embedding model. In *NIPS*.

Spandana Gella, Rico Sennrich, Frank Keller, and Mirella Lapata. 2017. Image pivoting for learning multilingual multimodal representations .

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Proceedings of NIPS*.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D* 42:335–346.

Z. Harris. 1954. Distributional Structure. *Word* 10(23):146—162.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision (CVPR)*. pages 1026–1034.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pages 770–778.

Felix Hill, Kyunghyun Cho, and Anna Korhonen. 2016a. Learning distributed representations of sentences from unlabelled data. In *Proceedings of NAACL*.

Felix Hill, Roi Reichart, and Anna Korhonen. 2016b. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics* .

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of SIGKDD*. pages 168–177.

Susan S Jones, Linda B Smith, and Barbara Landau. 1991. Object properties and knowledge in early lexical learning. *Child development* 62(3):499–516.

A. Joulin, L.J.P. van der Maaten, A. Jabri, and N. Vasilache. 2016. Learning visual features from large weakly supervised data. In *ECCV*.

A. Karpathy, A. Joulin, and L. Fei-Fei. 2014. Deep fragment embeddings for bidirectional image sentence mapping. In *Proceedings of NIPS*.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pages 3128–3137.

Douwe Kiela. 2017. Deep embodiment: grounding semantics in perceptual modalities (PhD thesis). Technical Report UCAM-CL-TR-899, University of Cambridge, Computer Laboratory.

Douwe Kiela and Léon Bottou. 2014. Learning Image Embeddings using Convolutional Neural Networks for Improved Multi-Modal Semantics. In *Proceedings of EMNLP*.

Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Ryan Kiros, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2015. Skip-thought vectors. In *Proceedings of NIPS*.

Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. 2015. Associating neural word embeddings with deep image representations using fisher vectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. pages 4437–4446.

Quoc V. Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of ICML*.

Jason Lee, Kyunghyun Cho, Jason Weston, and Douwe Kiela. 2017. Emergent translation in multi-agent communication. *CoRR* abs/1710.06922.

A. Li, A. Jabri, A. Joulin, and L.J.P. van der Maaten. 2016. Learning visual n-grams from web data. In *arxiv*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *European Conference on Computer Vision (ECCV)*. Springer, pages 740–755.

Max M. Louwerse. 2008. Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science* 59(1):617–645.

Thang Luong, Richard Socher, and Christopher D Manning. 2013. Better word representations with recursive neural networks for morphology. In *Proceedings of CoNLL*. pages 104–113.

Elman Mansimov, Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov. 2016. Generating images from captions with attention. In *International Conference on Learning Representations (ICLR)*.

J. Mao, W. Xu, Y. Yang, J. Wang, and A.L. Yuille. 2015. Deep captioning with multimodal recurrent neural networks. In *Proceedings of ICLR*.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Reffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of LREC*.

Allan Paivio. 1990. *Mental representations: A dual coding approach*. Oxford University Press.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of ACL*. page 271.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of ACL*. pages 115–124.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of EMNLP*.

Charles A Perfetti. 1998. The limits of co-occurrence: Tools and theories in language research. *Discourse Processes* 25(2&3):363—377.

Nghia The Pham, German Kruszewski, Angeliki Lazaridou, and Marco Baroni. 2015. Jointly optimizing word representations for lexical and sentential tasks with the c-phrase model. In *Proceedings of ACL*.

S. Reed, Z. Akata, H. Lee, and B. Schiele. 2016. Learning deep representations of fine-grained visual descriptions. In *Proceedings of CVPR*.

Brian Riordan and Michael N Jones. 2011. Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science* 3(2):303–345.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. 2014. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations (ICLR)*.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, Christopher Potts, et al. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of EMNLP*. pages 1631—1642.

I. Sutskever, O. Vinyals, and QV. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of NIPS*.

Alfred Tarski. 1944. The semantic conception of truth: and the foundations of semantics. *Philosophy and phenomenological research* 4(3):341–376.

Peter D. Turney and Patrick Pantel. 2010. From Frequency to Meaning: vector space models of semantics. *Journal of Artifical Intelligence Research* 37(1):141–188.

Aaron van den Oord, Nal Kalchbrenner, Lasse Espeholt, koray kavukcuoglu, Oriol Vinyals, and Alex Graves. 2016. Conditional image generation with pixelcnn decoders. In *Proceedings of NIPS*.

Ivan Vendrov, Ryan Kiros, Sanja Fidler, and Raquel Urtasun. 2015. Order-embeddings of images and language. In *International Conference on Learning Representations (ICLR)*.

O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of CVPR*.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation* 39(2):165–210.

John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. 2016. Towards universal paraphrastic sentence embeddings. In *International Conference on Learning Representations (ICLR)*.