

# Learning Zero-Shot Multifaceted Visually Grounded Word Embeddings via Multi-Task Training

Hassan Shahmohammadi Hendrik P. A. Lensch R. Harald Baayen

University of Tübingen

hassan.shahmohammadi@uni-tuebingen.de

{hendrik.lensch, harald.baayen}@uni-tuebingen.de

## Abstract

Language grounding aims at linking the symbolic representation of language (e.g., words) into the rich perceptual knowledge of the outside world. The general approach is to embed both textual and visual information into a common space -the grounded space- confined by an explicit relationship. We argue that since concrete and abstract words are processed differently in the brain, such approaches sacrifice the abstract knowledge obtained from textual statistics in the process of acquiring perceptual information. The focus of this paper is to solve this issue by implicitly grounding the word embeddings. Rather than learning two mappings into a joint space, our approach integrates modalities by implicit alignment. This is achieved by learning a reversible mapping between the textual and the grounded space by means of multi-task training. Intrinsic and extrinsic evaluations show that our way of visual grounding is highly beneficial for both abstract and concrete words. Our embeddings are correlated with human judgments and outperform previous works using pretrained word embeddings on a wide range of benchmarks. Our grounded embeddings are publicly available [here](#).

## 1 Introduction

The distributional hypothesis asserts that words occurring in similar contexts are semantically related (Harris, 1954). Current state-of-the-art word embedding models (Pennington et al., 2014; Peters et al., 2018a), despite their successful application to various NLP tasks (Wang et al., 2018), suffer from the lack of grounding in general knowledge (Harnad, 1990; Burgess, 2000), such as captured by human perceptual and motor systems (Pulvermüller, 2005; Theriault et al., 2009). To overcome this limitation, research has been directed to linking word embeddings to perceptual knowledge in visual scenes. Most studies have attempted to bring

visual and language representations into close vicinity in a common feature space (Silberer and Lapata, 2014; Kurach et al., 2017; Kiela et al., 2018).

However, studies of human cognition indicate that the brain processes abstract and concrete words differently (Paivio, 1990; Anderson et al., 2017) due to the difference in associated sensory perception. According to Montefinese (2019), similar activity for both categories are observed in the perirhinal cortex, a region related to memory and recognition, whereas in the parahippocampal cortex, associated with memory formation, higher activity only occurs for abstract words.

We argue that forcing the textual and visual modalities to be represented in a shared space causes grounded embeddings to suffer from the bias towards concrete words as reported by Park and Myaeng (2017); Kiela et al. (2018). Therefore, we propose a zero-shot approach that implicitly integrates perceptual knowledge into pre-trained textual embeddings (GloVe (Pennington et al., 2014) and fastText (Bojanowski et al., 2017)) via multi-task training. Our approach learns multifaceted grounded embeddings which capture multiple aspects of words' meaning and are highly beneficial for both concrete and abstract words.

Figure 1 lays out the architecture of our model. It learns a reversible mapping from pre-trained text-based embeddings to grounded embeddings which maintains the linguistic co-occurrence statistics while integrating visual information. The architecture features a similar structure as an auto-encoder (Press and Wolf, 2017) translating from words to grounded space and back. The training is carried out as multi-task learning by combining image captioning in two directions and image-sentence pair discrimination. At the core is a mapping matrix that acts as an intermediate representation between the grounded and textual space, which learns to visually ground the textual word vectors. This mapping is trained on a subset of words and then is

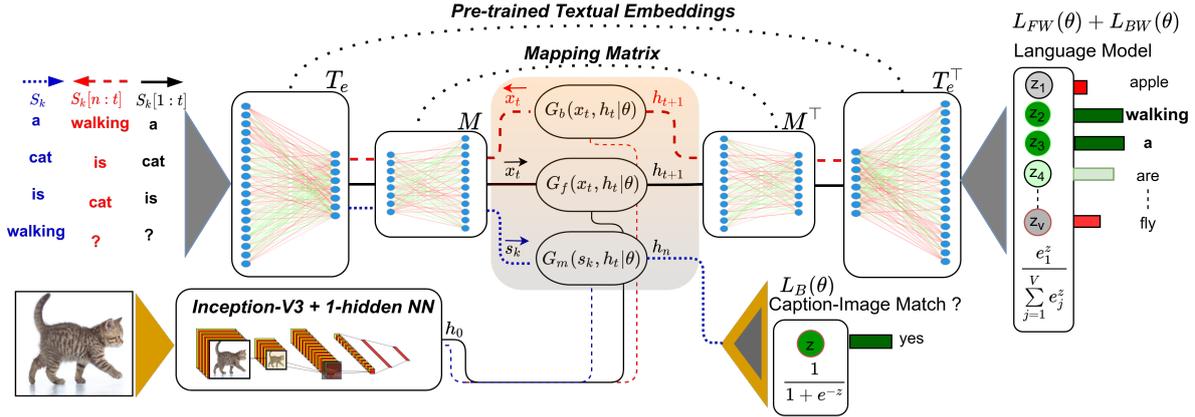


Figure 1: Our zero-shot model: 1. Two GRU based language-model tasks in forward ( $G_f$ ) and backward ( $G_b$ ) directions represented by solid black and dashed red lines. 2. A matching task predicting if the given (sentence, image) pair match (blue dotted line). The zero-shot mapping matrix  $M$ , shared by all the tasks, learns to visually ground the textual word vectors by learning a reversible mapping from textual space to grounded space.

applied to ground the full vocabulary of textual embeddings in a zero-shot manner.

We evaluate our grounded embeddings on both intrinsic and extrinsic tasks (Wang et al., 2019) and show that they outperform textual embeddings and previous related works in the majority of cases. Overall, our contributions are the following:

- we design a language grounding framework that can effectively ground different pre-trained word embeddings in a zero-shot manner;
- we create visually grounded versions of two popular word embeddings and make them publicly available;
- unlike many previous works, our embeddings support both concrete and abstract words;
- we show that visual grounding has the potential to refine the irregularities of a text-based vector space.

## 2 Related Works

The many attempts to combine images and text in order to obtain visually grounded word/sentence representations can be grouped into the following categories.

**Feature Level Fusion:** where the grounded embedding is the result of combining the visual and textual features. Combining strategies range from simple concatenation to adopting SVD and GRU gating mechanisms (Bruni et al., 2014; Kiela and Bottou, 2014; Kiros et al., 2018).

**Mapping to Perceptual Space:** this is usually a regression task predicting the image vector given its corresponding textual vector. The grounded embedding are extracted from an intermediate

layer in auto-encoders (Silberer and Lapata, 2014; Hasegawa et al., 2017), the output of an MLP (Collell Talleda et al., 2017) or an RNN (Kiela et al., 2018). Another method is mapping both modalities into a common space in which their distance is minimized (Kurach et al., 2017; Park and Myaeng, 2017).

**Equipping Distributional Semantic Models with Visual Context:** here images are treated as a context in the process of computing the word vectors. Many of these approaches modify the Word2Vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014) models by incorporating image features to the context for concrete words (Hill and Korhonen, 2014; Kottur et al., 2016; Zablocki et al., 2017; Ailem et al., 2018); minimizing the max-margin loss between the image-vector and its corresponding word vectors (Lazaridou et al., 2015); providing social cues based on child-directed speech along with visual scenes (Lazaridou et al., 2016); or by extracting the relationship between words and images using multi-view spectral graphs (Fukui et al., 2017).

**Hybrid:** this category covers the combination of previous methods and other strategies. Here, the grounded word vectors are usually the results of updating the textual word vectors during training (Mao et al., 2016) or the output of sentence encoders such as LSTM (Hochreiter and Schmidhuber, 1997). Such methods include predicting the image vector along with training a language model (Chrupała et al., 2015) or generating an alternative caption at the same time (Kiela et al., 2018). Other approaches such as using the coefficients of

classifiers for grounded representation have also emerged (Moro et al., 2019). Our model falls in the hybrid category as we take a multitasking approach. However, unlike some previous works (Kiela et al., 2018; Collell Talleda et al., 2017; Bordes et al., 2019) we do not impose explicit constraints between the image features and their captions. Our model learns the relationship indirectly via multi-task training.

### 3 Multi-Task Visual Grounding

In this section, we present the details of the developed method. The training data set  $D$  consists of image–caption pairs,  $(S_k, I_k) \in D$ , with  $S_k = [w_1, w_2 \dots w_n]$  being a sentence with  $n$  words describing the image  $I_k$ . We use the Microsoft\_COCO\_2017 dataset (Lin et al., 2014) in our experiments. Let  $T_e(w) \in \mathbb{R}^d$  be a pre-trained textual embedding of the word  $w$ , which has been trained on textual data only (e.g., GloVe). The objective is to train a mapping matrix  $M$  to ground the word vector  $T_e(w)$  visually, resulting in a grounded embedding  $G_e(w) = T_e(w) \cdot M$ , where  $G_e(w) \in \mathbb{R}^c$ . To do so, we train the matrix  $M$  to refine the textual vector space via two image-based language model tasks and a binary discrimination task on image-sentence pairs. For the language models, a GRU (Cho et al., 2014) is trained to predict the next word, given the previous words in the sentence provided as image caption, and its associated image vector. The transpose of the textual embedding  $T_e$  is used to compute the probability distribution over the vocabulary (see Figure 1). We employ an identical scenario to form a second language model task using another GRU, where the sentence is fed backward into the model.

The image-sentence discrimination is a binary classification task predicting whether the given sentence  $S_k$  represented in the grounded space matches the image  $I_k$ . By training the model simultaneously on these three tasks confined by a linear transformation, we augment the visual information into the grounded embeddings (output of mapping matrix in Figure 1) while preserving the underlying structure of the textual embeddings.

#### 3.1 Language Model

Given the input caption associated with image  $I_k$  as  $S_k = [w_1, w_2 \dots w_n]$ , we first encode the words using a pre-trained textual embedding  $T_e$  to obtain the embeddings as  $S_t = [t_1, t_2 \dots t_n]$ . We then linearly

project these embeddings from the textual space into the visually grounded space via the trainable mapping matrix  $M$  as  $G_e(S_k) = S_t \cdot M$ , to obtain a series of grounded vectors  $G_e(S_k) = [x_1, x_2 \dots x_n]$  where  $x_i \in \mathbb{R}^c$ . In the grounded space, the perceptual information of the image  $I_k$  corresponding to  $S_k$  is fused using a single-layer GRU ( $G_f$  ( $f$ -forward) in Figure 1) that predicts the next output  $h_{t+1} = GRU_f(x_t, h_t | \theta)$ , where  $\theta$  denotes the trainable parameters,  $x_t$  the current input ( $G_e(w_t)$ ), and  $h_t \in \mathbb{R}^c$  the current hidden state.

Image information is included by initializing the first hidden state  $h_0$  with the image vector of  $I_k$ . The GRU update gate propagates perceptual knowledge from images into the mapping matrix. This has been shown to be more effective than providing the image vector at each time step as input (Mao et al., 2016).

The transpose of the mapping matrix ( $M^\top$ ) is used to map back from grounded space to the textual space. That is, the output of the GRU in each time-step is mapped back into the textual space as  $w_{next} = h_t \cdot M^\top$ , where  $w_{next} \in \mathbb{R}^d$  is an approximation of the next word’s textual embedding. The mapping matrix  $M$  is used to both encode and decode into/from the grounded space. This improves generalization (Press and Wolf, 2017) and prevents the vanishing gradient problem compared to the case where the mapping matrix is only used at the beginning of the network (Mao et al., 2016).  $w_{next}$  is fed into the transpose of the textual embeddings in the same scenario:  $z = T_e^\top(w_{next})$ , where  $z \in \mathbb{R}^{|V|}$  and  $V$  indicates the vocabulary. The final probability distribution over  $V$  is computed by a softmax:

$$P(y = j | z) = \frac{e^{z^\top W_j}}{\sum_{c=1}^V e^{z^\top W_c}} \quad (1)$$

Defining the input (previous words and the image vector) and the predicted output (next word prediction) as above, we minimize the categorical cross entropy which is computed for batch  $B$  as:

$$\mathcal{L}_{FW}(\theta) = -\frac{1}{|B|} \sum_{i \in B} \sum_{c \in V} y_{i,c} \log(\hat{y}_{i,c}), \quad (2)$$

Where  $\hat{y}_{i,c}$  and  $y_{i,c}$  are the predicted probability and ground truth for sample  $i$  with respect to the class  $c$ .

Moreover, we define a second similar task: Given the input caption associated with image  $I_k$  as  $S_k = [w_1, w_2 \dots w_n]$ , we reverse the order of

the words:  $S_k = [w_n, w_{n-1} \dots w_1]$  and use another GRU ( $G_b$  ( $b$ -backward) in Figure 1) with identical structure trained on the loss  $\mathcal{L}_{BW}(\theta)$ . The rest of the network is shared between these two tasks. Having this backward language model is analogous to bi-directional GRUs (Schuster and Paliwal, 1997) which, however, can not be used directly since the ground truth would be exposed by operating in both directions.

### 3.2 Image-sentence discrimination

Even though context-driven word representations are a powerful way to obtain word embeddings (Pennington et al., 2014; Peters et al., 2018a), the performance of such models varies on language-  
vision tasks (Burns et al., 2019). Therefore, we propose yet another task to align the textual word vectors to their real-world relations in the images. The discrimination task predicts whether the given image and sentence describe the same content or not (shown by ‘caption-image match?’ in Figure 1). These types of tasks have been shown effective for learning cross-modality representations (Lu et al., 2019; Tan and Bansal, 2019).

Given the input caption for image  $I_k$  as  $S_k = [w_1, w_2 \dots w_n]$ , after projecting the embeddings into the grounded space as before, we encode the whole sentence by employing a third single-layer GRU ( $G_m$  in Figure 1) with the same structure as before  $h_n = GRU_m(G_e(S_k), h_0|\theta)$ . Where the last output  $h_n$  encodes the whole sentence.  $h_0$  is again initialized with the image vector of  $I_k$ . The final output is computed by a sigmoid function. This task shares the mapping matrix  $M$  and the textual embeddings  $T_e$ . We minimize the binary cross entropy, which could be computed for each batch as:

$$\mathcal{L}_B(\theta) = -\frac{1}{|B|} \sum_{i \in B} y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \quad (3)$$

For negative mining, half of the captions in each batch are replaced with captions of different, random images.

### 3.3 Regularization and overall loss

All the three tasks explained above share the pre-trained textual embeddings (see Figure 1) which gives rise to the question of whether the textual embeddings should be updated or kept fixed during training. By updating, we might distort the pre-trained semantic relations, especially given our

limited training data. Keeping them fixed, on the other hand, does not provide the flexibility to generate the desired grounding as these embeddings are noisy and not perfect (Yu et al., 2017). To prevent distorting the semantic information of words while retaining sufficient flexibility, we propose the following regularization on the embedding matrix  $T_e$ :

$$\mathcal{R}(\alpha, \beta) = \frac{\alpha}{|V|} \sum_{w \in V} \left| \beta - \frac{w_e \cdot w_n}{\|w_e\| \|w_n\|} \right|, \quad (4)$$

where  $\alpha$  controls the overall impact and  $\beta$  controls how much the new word vectors  $w_n$  are allowed to deviate from the pre-trained embedding  $w_e$ .  $\beta = 1$  indicates no deviation and  $\beta = 0$  allows for up to 90 degree deviation from  $w_n$  when minimizing the equation. We join all the tasks into a single model and minimize the following loss:

$$\mathcal{L}_{All}(\Theta) = \mathcal{L}_{FW}(\theta) + \mathcal{L}_{BW}(\theta) + \mathcal{L}_B(\theta) + \mathcal{R}(\alpha, \beta) \quad (5)$$

where  $\Theta$  denotes all the trainable parameters.

## 4 Experimental setup

We use the Microsoft\_COCO\_2017 dataset (Lin et al., 2014) for training. Each sample contains an image with 5 captions. The dataset is split into 118k train and 5k validation samples. Each batch includes 256 image vectors along with one of their captions. Hence, multiple image vectors might occur in each batch. Image vectors are obtained by transferring the penultimate layer of pre-trained Inception-V3 (Szegedy et al., 2016) trained on ImageNet (Deng et al., 2009). A NN with one hidden layer and  $\tanh$  activation is employed to project the image vectors into the initial hidden state of the GRUs:  $h_t \in \mathbb{R}^{1024}$ . We lowercase all the words, delete the punctuation marks, and only keep the top 10k most frequent words. Two popular pre-trained textual word embeddings namely GloVe (*crawl-300d-2.2M-cased*) and fastText (*crawl-300d-2M-SubW*) are used for initialization of the embedding  $T_e$ . The mapping matrix  $M$  transforms the textual embeddings into the grounded space. We investigate the best dimension of this step and the improvement over pure textual embeddings in the next sections. Batch normalization (Ioffe and Szegedy, 2015) is applied after each GRU. For the regularization,  $\mathcal{R}(\alpha = 0.001, \beta = 1)$  for GloVe and  $\mathcal{R}(\alpha = 0.01, \beta = 0)$  for fastText yielded the best relative results by meta parameter search. This shows that FastText embeddings

require more deviation ( $\beta = 0$  indicates 90 degree deviation) to adapt to the proposed tasks. We trained the model for 20 epochs with 5 epochs tolerance early stopping using NAdam (Dozat, 2016) with a learning rate of 0.001.

As we train a single mapping matrix  $M$  for projecting from textual to grounded space, it can be used after the training to transfer out-of-vocabulary (OoV) word-vectors into the grounded space in a zero-shot manner. This way, visually grounded versions of both Glove and fastText are obtained despite being exposed to only  $10k$  words.

## 5 Evaluations

While the question of what is a good word embedding model is still open (Wang et al., 2019), there are two main categories of evaluation methods: intrinsic and extrinsic. Intrinsic evaluators measure the quality of word embeddings independent of any downstream tasks. For instance, quality can be assessed by comparing similarities between embeddings with word similarities as perceived by human raters. Extrinsic evaluators on the other hand assess the performance based on sentence-level downstream tasks. There is not necessarily a positive correlation between intrinsic and extrinsic methods for a word embedding model (Wang et al., 2019). Nonetheless, we use both types of evaluators to compare our visually grounded embeddings with those presented in related works as well as to purely text-based embeddings.

**Baselines:** we considered two types of embeddings as baselines 1) the pre-trained textual embeddings  $T_e$ , 2)  $T_e$  refined based only on the captions without injecting any image information using a similar language modeling task  $\mathcal{L}_{FW}$  with a one-layer GRU ( $h_t \in \mathbb{R}^{1024}$ ) followed by a fully connected layer. We refer to the second baseline as C\_GloVe and C\_fastText for Glove and fastText trained only on captions.

**Intrinsic Evaluators:** We evaluate on some of the common lexical semantic similarity benchmarks: MEN (Bruni et al., 2014), SimLex999 (Hill et al., 2015), Rare-Words (Luong et al., 2013), MTurk771 (Halawi et al., 2012), WordSim353 (Finkelstein et al., 2001), and SimVerb3500 (Gerz et al., 2016). The evaluation metric is the Spearman correlation between the predicted cosine similarity vector and the ground truth.

**Extrinsic Evaluators:** We evaluate on the semantic textual similarity benchmarks (STS) from year 2012 to 2016 using SentEval (Conneau and Kiela, 2018). Here, the task is to measure the semantic equivalence of a pair of sentences solely based on their cosine coefficient. We are particularly interested in these benchmarks for two reasons. 1) They evaluate the generalization power of the given vector space without any fine-tuning. 2) Since they contain sentences from various sources such as news headlines and public forums, they reveal whether abstract knowledge is still preserved by our framework. We used BoW (averaging) to obtain sentence representations. While BoW is a simple sentence encoder, it is a great tool to evaluate the underlying structure of a vector space. For instance, the BoW representation of a pair of sentences such as ‘her dog is very smart’ and ‘his cat is too dumb’ are, unfortunately, very similar in a vector space that does not distinguish dissimilar from related words (e.g., smart and dumb). We will show that our model properly refines the textual vector space and alleviates these kinds of irregularities.

Model	RW	MEN	WSim 353	MTurk 771	SimVerb 3500	SimLex 999	Mean
GloVe	45.5	80.5	73.8	71.5	28.3	40.8	56.7
C_GloVe	46	82.1	74.1	72.3	29.3	43.3	57.85
VGE_G	<b>52.6</b>	<b>85.1</b>	<b>78.9</b>	<b>73.4</b>	<b>37.4</b>	<b>51.8</b>	<b>63.2</b>
FastText	56.1	81.5	72.2	75.1	37.8	47.1	61.6
C_fastText	49.2	68.3	58.1	56.8	30.3	41.9	50.76
VGE_F	<b>57.8</b>	<b>83.6</b>	<b>73.9</b>	<b>76.1</b>	<b>39.2</b>	<b>49.0</b>	<b>63.2</b>

Table 1: Intrinsic evaluation. Visual grounding (denoted by ‘VGE’) improves the results compared to the baselines on all test sets.

Model	RW	MEN	WSim 353	MTurk 771	SimVerb 3500	SimLex 999
VGE_G	52.6	<b>85.1</b>	<b>78.9</b>	73.4	37.4	51.8
VGE_F	<b>57.8</b>	83.6	73.9	<b>76.1</b>	<b>39.2</b>	49.0
Cap2Both	48.7	81.9	71.2	–	–	46.7
Cap2Img	52.3	84.5	75.3	–	–	51.5
Park et al.	–	83.8	77.5	–	–	<b>58.0</b>
Park_VG.	–	–	–	–	–	15.7
Collell et al.	–	81.3	–	–	28.6	41.0

Table 2: Comparison of grounded embeddings to previous work on intrinsic tasks. Ours are denoted by VGE.

## 6 Results

**Intrinsic Evaluation – Baselines:** Table 1 shows the intrinsic evaluation results for the baselines and our visually grounded embeddings (VGE\_F and VGE\_G for visually grounded fastText and Glove respectively). In general, fastText performs better on word-level tasks compared to GloVe, prob-

ably because it provides more context for each word by leveraging from its sub-words. The results also validate the efficacy of our proposed model since updating the embeddings on captions alone (C\_fastText and C\_GloVe) brings subtle or no improvements. By the proposed visual grounding, significant improvements are achieved on *all* datasets for both fastText and GloVe. Analyzing why the improvement varies across different datasets is difficult. However, the table reveals interesting properties. For instance, the improvement on SimLex999, which focuses more on the similarity between words, is larger than that on WSim353, which does not distinguish between similarity and relatedness. Hence, visual grounding seems to prioritize similarity over relatedness. Considering the overall performance, it enhances both embeddings to the same level despite their fundamental differences.

#### **Intrinsic Evaluation – Grounded Embeddings:**

We compare our model to related grounded embeddings by (Collell Talleda et al., 2017; Park and Myaeng, 2017; Kiros et al., 2018; Kiela et al., 2018) (Table 2). We limit our comparison to those who adopted the pre-trained GloVe or fastText since these pre-trained models alone outperform many visually grounded embeddings such as (Hasegawa et al., 2017; Zablocki et al., 2017) on many of our evaluation datasets.

Conceptually, Kiela et al. (2018) also induces visual grounding on GloVe by using the MSCOCO data set. Even though they propose a number of tasks for training (Cap2Img: predicting the image vector from its caption, Cap2Cap: generate an alternative caption of the same image; Cap2Both: training by Cap2Cap and Cap2Img simultaneously) our model clearly outperforms them as ours integrate visual information without degraded performance on abstract words.

Park and Myaeng (2017) proposed a polymodal approach by creating and combining six different types of embeddings (linear and syntactic contexts, cognition, sentiment, emotion, and perception) for each word. Even though they used two pre-trained embeddings (GloVe and Word2vec) and other resources, our model still outperforms their approach on MEN and WSim353, but their approach is better on Simlex999. This performance can be attributed to the many-modality training as using only their visually grounded embeddings (Park\_VG) performs much worse. This clearly shows that their visual

embeddings do not benefit abstract words (cf. Park and Myaeng, 2017). In summary, our approach benefits from capturing different perspectives of the words’ meanings by learning the reversible mapping in the context of multi-task learning.

**Fine-Grained Intrinsic Evaluation:** we further evaluate our model on the different categories of SimLex999 divided into nine sections: all (the whole dataset), adjectives, nouns, verbs, concreteness quartiles (from 1 to 4 increasing the degree of concreteness), and hard pairs. The hard section indicates 333 pairs whose similarity is hard to discriminate from relatedness. The results for our best embeddings on SimLex999 (VGE\_G) are shown in Table 3. We see a large improvement over GloVe in all categories. Some previous approaches such as (Park and Myaeng, 2017) concluded that perceptual information would be beneficial only to concrete words (e.g., apple, table) and would adversely affect abstract words (e.g., happy, freedom). However, our model succeeds in maintaining high-precision co-occurrence statistics from the textual model while augmenting these with perceptual information, in such a way that the representations for abstract words are actually enhanced. Therefore, it outperforms GloVe not only on concrete pairs (conc-q4) but also on highly abstract pairs (conc-q1).

We compared the results on SimLex999 with another recent visually grounded model called Picturebook (Kiros et al., 2018), which employs a multi-modal gating mechanism (similar to a LSTM and GRU update gate) to fuse the GloVe and Picturebook embeddings (Table 3). It uses image feature vectors pre-trained on a fine-grained similarity task with 100+ million images (Wang et al., 2014). Picturebook’s performance is highly biased toward concrete words (conc-q3, conc-q4) and performs worse than GloVe by nearly 29% on highly abstract words (conc-q1). Picturebook + GloVe on the other hand shows better results but still performs worse on highly abstract words and adjectives. Our model (VGE\_G) can generalize across different categories and outperforms Picturebook+Glove with a large margin on most of the categories while being quite comparable on the others.

**Refining the Textual Vector Space:** Our grounded embeddings, while improving relatedness scores, prioritize similarity over relatedness. This is further demonstrated through inspection of nearest neighbors (Table 5). Given the word ‘bird’,

Model	All	Adjs	Nouns	Verbs	Conc-q1	Conc-q2	Conc-q3	Conc-q4	Hard
GloVe	40.8	62.2	42.8	19.6	43.3	41.6	42.3	40.2	27.2
VGE_G (ours)	<b>51.8</b>	<b>72.1</b>	52.0	<b>35</b>	<b>53.1</b>	<b>54.8</b>	47.4	56.8	<b>38.3</b>
Picturebook	37.3	11.7	48.2	17.3	14.4	27.5	46.2	<b>60.7</b>	28.8
Picturebook+GloVe	45.5	46.2	<b>52.1</b>	22.8	36.7	41.7	<b>50.4</b>	57.3	32.5

Table 3: SimLex999 (Spearman’s  $\rho$ ) results. Conc-q1 and Conc-q4 contain the most abstract and concrete words respectively. Our embeddings (VGE\_G) generalize across different word types and strongly outperform all the others on most of the categories.

Model	STS12	STS13	STS14	STS15	STS16	Mean
GloVe	52.25	49.59	54.72	56.25	51.39	52.84
C_GloVe	53.27	50.56	56.72	57.86	52.11	54.10
VGE_G	<b>55.31</b>	<b>57.24</b>	<b>65.54</b>	<b>67.61</b>	<b>65.87</b>	<b>62.35</b>
Fasttext	22.95	24.63	31.37	37.71	29.34	29.2
C_fasttext	29.69	23.80	37.58	45.29	29.34	33.14
VGE_F	<b>31.78</b>	<b>32.26</b>	<b>42.51</b>	<b>48.79</b>	<b>38.15</b>	<b>38.70</b>
VGE_G (ours)	<b>55</b>	<b>57</b>	<b>66</b>	<b>68</b>	66	<b>62.40</b>
Word2vec	52	<b>58</b>	<b>66</b>	<b>68</b>	65	61.80
ELMo (top_layer)	54	49	62	67	63	59.00
ELMo (all_layers)	<b>55</b>	51	63	69	64	60.40
Power-mean	54	52	63	66	<b>67</b>	60.40

Table 4: Comparison (Pearson correlation  $\times 100$ ) of our embeddings (VGE\_\*) with baselines (first two sections) and other word embeddings (bottom) on STS.

GloVe returns ‘turtle’ and ‘nest’ while grounded GloVe returns ‘sparrow’ and ‘avian’, which both reference birds. Moreover, our embeddings retrieve more meaningful words regardless of the degree of abstractness. For the word ‘happy’ for example, GloVe suffers from a bias toward dissimilar words with high co-occurrence such as ‘everyone’, ‘always’, and ‘wish’. This issue is intrinsic to the fundamental assumption of the distributional hypothesis that words in the same context tend to be semantically related. Therefore, GloVe embeddings, even though trained on 840 billion tokens, still reports antonyms such as ‘smart’ and ‘dumb’ as very similar. In addition, common misspellings of words (e.g., ‘together’) while serving the same role, occur with different frequencies in changing context. Hence, they are pulled apart in purely text-based vector spaces. However, our visual grounding model clearly puts them in the same cluster. Our model therefore seems to refine the text-based vector space by aligning it (via the mapping matrix) with real-world relations (in the images). This refinement generalizes to all the words by using our zero-shot mapping matrix which explains the improvement on highly abstract words. A sample of nearest neighbors for FastText and VGE\_F is available in Appendix B. However, since FastText already performs quite well on intrinsic tasks, the difference with its grounded version is subtle which also confirms the results in Table 1.

**Extrinsic Evaluation:** Table 4 shows the results on semantic similarity benchmarks. Both grounded embeddings strongly outperform their textual version on *all* benchmarks. While fastText outperforms GloVe on intrinsic tasks, GloVe is superior here. The reason might be that unlike fastText GloVe treats each word as a single unit and takes into account the global co-occurrences of words. This probably helps to capture the high-level structure of words (e.g., in sentences). Considering the mean score, our model boosts both embeddings approximately by 10 percent.

Furthermore, while we are well aware that our simple averaging model cannot compete with the state-of-the-art sequence models (Gao et al., 2021) on the sentence level STS task, we compare it to other word embeddings to highlight the contribution of visual grounding. Table 4 (bottom) shows the results of our best model (VGE\_G) with other textual word embeddings namely ELMo (Peters et al., 2018b), Word2Vec (Mikolov et al., 2013), and Power-Mean (Rücklé et al., 2018) reported by (Perone et al., 2018). While the textual GloVe is the second-worst model (by mean score: 52.84) in the table, its grounded version VGE\_G is the best one. Overall, these results confirm that 1) our grounding framework effectively integrates perceptual knowledge that is missing in purely text-based embeddings and 2) visual grounding is highly beneficial for downstream language tasks. It would be interesting to see if our findings extend to grounded sentence embedding models (Sileo, 2021; Bordes et al., 2019; Tan and Bansal, 2020) for instance by training transformer-based models such as BERT (Devlin et al., 2018) on top of our embeddings. However, we postpone this to the future since our focus here is on grounding word embeddings.

## 7 Model Analysis

We further analyze the performance of our model from different perspectives as follows.

**Dependency on the Encoding Dimension c:** We train our model with different dimensions of the

happy		sad		big		bird		horse		together		smart	
G	V	G	V	G	V	G	V	G	V	G	V	G	V
lucky	pleased	sadly	saddened	hard	humongous	turtle	sparrow	dog	racehorse	well	together	sensible	witty
everyone	delighted	shame	tragic	little	Big	nest	Birds	riding	Thoroughbred	bring	togheter	dumb	shrewd
love	merry	horrible	mournful			squirrel	avian	ponies	Horses	both	together	sophisticated	intelligent
always	thrilled	scared	saddening					donkey	steed	they	together	attractive	resourceful
wish	joyful	awful	sorrowful							apart	together	wise	quick-witted
hope	hapy	pity	Sad							up	2gether		
		kinda	heartbreaking							them	together		
		sorry	heartbroken							put	together		
										along	toghether		
										with	gether		

Table 5: Results of 10 nearest neighbors for Glove (G) and VGE\_G (V). Only the differing neighbors are reported. While GloVe retrieves more related words, ours (VGE\_G) focuses on similar words. Overall, VGE\_G is closer to human judgment and retrieves highly semantically similar words.

Dataset	Best $\alpha$	Acc. with best $\alpha$	Acc. with $\alpha = 1$
RareWords	1.00	52.6	52.6
MEN	0.63	85.2	85.1
WSim353	0.57	79.3	78.9
Mturk771	0.52	74.2	73.4
SimVerb3500	1.00	37.4	37.4
SimLex999	1.00	51.8	51.8

Table 6: Sensitivity analysis (Spearman’s  $\rho$ ) on intrinsic datasets.  $\alpha = 1$  indicates no use of GloVe and  $\alpha = 0$  means no use of VGE\_G. Pure grounded embeddings alone yield the best results on 3 of the datasets.

Embeddings	$\mathcal{L}_{FW}$	$\mathcal{L}_{FW} + \mathcal{L}_{BW}$	$\mathcal{L}_{FW} + \mathcal{L}_{BW} + \mathcal{L}_B$	$\mathcal{L}_{All} + \mathcal{R}(\alpha, \beta)$
VGE_G	61.60	61.82	62.66	63.20
VGE_F	61.70	61.83	61.60	63.20

Table 7: Mean score (Spearman’s  $\rho$ ) on intrinsic datasets with respect to each task.  $\mathcal{L}_{All}$  refers to all the three tasks and  $\mathcal{R}(\alpha, \beta)$  the regularization loss.

grounded embeddings and measure the mean accuracy of all the intrinsic datasets. Table 8 shows the results using GloVe and VGE\_G with different sizes. Significant improvement is already achieved keeping the original dimension of GloVe (300). Higher dimensions up to a certain threshold (1024) increase the accuracy but beyond this point, the model starts to overfit.

**Dependency on the Textual Embeddings:** Further, we analyze how much of GloVe’s original properties are maintained by the visual grounding. Given  $V_w$  and  $G_w$  as the VGE\_G and GloVe vectors for the word  $w$ , we create a vector containing both embeddings  $E_w = [(1 - \alpha)G_w; \alpha V_w]$ . Varying the relative weight  $\alpha \in (0, 1]$  we evaluate on the intrinsic datasets in Table 6. Three of the datasets yield the best results using only the grounded embeddings. The reduction in accuracy regarding ‘MEN’ is also very subtle. On ‘WSim353’ and ‘Mturk771’, however, the best results are achieved with  $\alpha \approx 0.5$ . This might be

Model dimensions	G_300	V_300	V_512	V_800	V_1024	V_2048
Mean Score	56.7	62.4	62.6	63.1	<b>63.2</b>	62.5

Table 8: Effect of grounded word-vectors magnitude on intrinsic tasks. ‘G’ and ‘V’ refers to Glove and VGE\_G respectively. Significant improvement is achieved even with the same size as the textual GloVe.

because these datasets focus on the relatedness of words while SimLex999 for instance distinguishes between similarity and relatedness.

**Ablation Study:** We further analyze the contribution of each task by performing an ablation evaluation. Table 7 shows the mean score on all the intrinsic datasets (see Table 1) with respect to each loss for both embeddings. While both GloVe and FastText show the same behaviour for language model tasks, fastText embeddings require more deviation ( $\beta = 0$  in  $\mathcal{R}(\alpha, \beta)$ ) to adapt to the binary discrimination task ( $\mathcal{L}_{BW}$ ). Textual embeddings  $T_e$  were frozen for all the cases except for  $\mathcal{L}_{All}$ . Even though the best performance, considering all the datasets, is achieved by using all the losses (including the regularization), each loss contributes differently to the overall performance. A more detailed ablation study based on the SimLex999 dataset is provided in Appendix A.

**Connections to Human Cognition:** Motivated by the different processing patterns of abstract and concrete words in the brain Montefinese (2019), we showed that it is possible to benefit from visual information without learning the two modalities in a joined space. Our experiments show that leveraging visual knowledge to inform the distributional models about the real world might be a better way of integrating language and vision. These modalities while separated could be informed and aligned with each other.

## 8 Conclusion

We investigated the effect of integrating perceptual knowledge from images into word embeddings via multi-task training. We constructed the visually grounded versions of GloVe and fastText by learning a zero-shot transformation from textual to grounded space trained on the MSCOCO dataset. Results on intrinsic and extrinsic evaluation support that visual grounding benefits current textual word embedding models. The major findings in our experiments are as follows:

- a) Our improvement of visual grounding is not limited to words with concrete meanings and covers highly abstract words as well.
- b) Discrimination between relatedness and similarity is more precise when using grounded embeddings.
- c) Perceptual knowledge can profitably be transferred to purely textual downstream tasks.

Moreover, we showed that visual grounding has the potential to refine the irregularities in textual vector spaces by aligning words with their real-world relations. This paves the way for future research on how visual grounding could resolve the problem of dissimilar words that occur frequently in the same context (e.g., small and big). In the future, we will investigate whether transformer blocks could profitably replace the GRU cells since they lead the state-of-the-art in many downstream sentence tasks. Moreover, while thus far our focus has been on words, a similar approach could be extended to obtain grounded sentence representations.

## Acknowledgements

This work has been supported by EXC number 2064/1 – Project number 390727645, as well as by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Hassan Shahmohammadi. The third author was supported by ERC-WIDE (European Research Council – Wide Incremental learning with Discrimination nEtworks), grant number 742545.

## References

- Melissa Ailem, Bowen Zhang, Aurelien Bellet, Pascal Denis, and Fei Sha. 2018. [A probabilistic model for joint learning of word embeddings from texts and images](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1478–1487, Brussels, Belgium. Association for Computational Linguistics.
- 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*, 5:17–30.
- Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. [Enriching word vectors with subword information](#). *Transactions of the Association for Computational Linguistics*, 5:135–146.
- Patrick Bordes, Eloi Zablocki, Laure Soulier, Benjamin Piwowarski, and Patrick Gallinari. 2019. [Incorporating visual semantics into sentence representations within a grounded space](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 696–707, Hong Kong, China. Association for Computational Linguistics.
- Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47.
- Curt Burgess. 2000. Theory and operational definitions in computational memory models: A response to glenbergh and robertson. *Journal of Memory and Language*, 43(3):402–408.
- Andrea Burns, Reuben Tan, Kate Saenko, Stan Sclaroff, and Bryan A Plummer. 2019. Language features matter: Effective language representations for vision-language tasks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7474–7483.
- Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.
- Grzegorz Chrupała, 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 (Volume 2: Short Papers)*, pages 112–118, Beijing, China. Association for Computational Linguistics.

- Guillem Collell Talleda, 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)*, pages 4378–4384. 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. European Language Resources Association (ELRA).
- Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Timothy Dozat. 2016. Incorporating nesterov momentum into adam.
- 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*, pages 406–414.
- Kazuki Fukui, Takamasa Oshikiri, and Hidetoshi Shimodaira. 2017. [Spectral graph-based method of multimodal word embedding](#). In *Proceedings of TextGraphs-11: the Workshop on Graph-based Methods for Natural Language Processing*, pages 39–44, Vancouver, Canada. Association for Computational Linguistics.
- Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.
- Daniela Gerz, Ivan Vulić, Felix Hill, Roi Reichart, and Anna Korhonen. 2016. [SimVerb-3500: A large-scale evaluation set of verb similarity](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2173–2182, Austin, Texas. Association for Computational Linguistics.
- Guy Halawi, Gideon Dror, Evgeniy Gabrilovich, and Yehuda Koren. 2012. Large-scale learning of word relatedness with constraints. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1406–1414.
- Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346.
- Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.
- Mika Hasegawa, Tetsunori Kobayashi, and Yoshihiko Hayashi. 2017. [Incorporating visual features into word embeddings: A bimodal autoencoder-based approach](#). In *IWCS 2017 — 12th International Conference on Computational Semantics — Short papers*.
- 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)*, pages 255–265, Doha, Qatar. Association for Computational Linguistics.
- Felix Hill, Roi Reichart, and Anna Korhonen. 2015. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*.
- Douwe Kiela and Léon Bottou. 2014. [Learning image embeddings using convolutional neural networks for improved multi-modal semantics](#). In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 36–45, Doha, Qatar. Association for Computational Linguistics.
- 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, Volume 1 (Long Papers)*, pages 408–418, New Orleans, Louisiana. Association for Computational Linguistics.
- Jamie Kiros, William Chan, and Geoffrey Hinton. 2018. [Illustrative language understanding: Large-scale visual grounding with image search](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 922–933, Melbourne, Australia. Association for Computational Linguistics.
- Satwik Kottur, Ramakrishna Vedantam, José MF Moura, and Devi Parikh. 2016. Visual word2vec (vis-w2v): Learning visually grounded word embeddings using abstract scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4985–4994.
- Karol Kurach, Sylvain Gelly, Michal Jastrzebski, Philip Haeusser, Olivier Teytaud, Damien Vincent,

- and Olivier Bousquet. 2017. Better text understanding through image-to-text transfer. *arXiv preprint arXiv:1705.08386*.
- Angeliki Lazaridou, Grzegorz Chrupała, Raquel Fernández, and Marco Baroni. 2016. **Multimodal semantic learning from child-directed input**. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 387–392, San Diego, California. Association for Computational Linguistics.
- Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015. **Combining language and vision with a multimodal skip-gram model**. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 153–163, Denver, Colorado. Association for Computational Linguistics.
- 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*, pages 740–755. Springer.
- Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*.
- Thang Luong, Richard Socher, and Christopher Manning. 2013. **Better word representations with recursive neural networks for morphology**. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 104–113, Sofia, Bulgaria. Association for Computational Linguistics.
- Junhua Mao, Jiajing Xu, Kevin Jing, and Alan L Yuille. 2016. Training and evaluating multimodal word embeddings with large-scale web annotated images. In *Advances in neural information processing systems*, pages 442–450.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.
- Maria Montefinese. 2019. Semantic representation of abstract and concrete words: a minireview of neural evidence. *Journal of neurophysiology*, 121(5):1585–1587.
- Daniele Moro, Stacy Black, and Casey Kennington. 2019. Composing and embedding the words-as-classifiers model of grounded semantics. *arXiv preprint arXiv:1911.03283*.
- Allan Paivio. 1990. *Mental representations: A dual coding approach*. Oxford University Press.
- Joohee Park and Sung-hyon Myaeng. 2017. **A computational study on word meanings and their distributed representations via polymodal embedding**. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 214–223, Taipei, Taiwan. Asian Federation of Natural Language Processing.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. **GloVe: Global vectors for word representation**. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.
- Christian S Perone, Roberto Silveira, and Thomas S Paula. 2018. Evaluation of sentence embeddings in downstream and linguistic probing tasks. *arXiv preprint arXiv:1806.06259*.
- Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018a. **Deep contextualized word representations**. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.
- Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018b. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.
- Ofir Press and Lior Wolf. 2017. **Using the output embedding to improve language models**. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 157–163, Valencia, Spain. Association for Computational Linguistics.
- Friedemann Pulvermüller. 2005. Brain mechanisms linking language and action. *Nature reviews neuroscience*, 6(7):576–582.
- Andreas Rücklé, Steffen Eger, Maxime Peyrard, and Iryna Gurevych. 2018. Concatenated power mean word embeddings as universal cross-lingual sentence representations. *arXiv preprint arXiv:1803.01400*.
- Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.
- 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 (Volume 1: Long Papers)*, pages 721–732, Baltimore, Maryland. Association for Computational Linguistics.

- Damien Sileo. 2021. Visual grounding strategies for text-only natural language processing. *arXiv e-prints*, pages arXiv-2103.
- Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.
- Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.
- Hao Tan and Mohit Bansal. 2020. Vokenization: improving language understanding with contextualized, visual-grounded supervision. *arXiv preprint arXiv:2010.06775*.
- David J Theriault, Richard H Yaxley, and Rolf A Zwaan. 2009. The role of color diagnosticity in object recognition and representation. *Cognitive Processing*, 10(4):335.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. [GLUE: A multi-task benchmark and analysis platform for natural language understanding](#). In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.
- Bin Wang, Angela Wang, Fenxiao Chen, Yuncheng Wang, and C-C Jay Kuo. 2019. Evaluating word embedding models: Methods and experimental results. *APSIPA transactions on signal and information processing*, 8.
- Jiang Wang, Yang Song, Thomas Leung, Chuck Rosenberg, Jingbin Wang, James Philbin, Bo Chen, and Ying Wu. 2014. Learning fine-grained image similarity with deep ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1386–1393.
- Liang-Chih Yu, Jin Wang, K. Robert Lai, and Xuejie Zhang. 2017. [Refining word embeddings for sentiment analysis](#). In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 534–539, Copenhagen, Denmark. Association for Computational Linguistics.
- Eloi Zablocki, Benjamin Piwowarski, Laure Soulier, and Patrick Gallinari. 2017. Learning multi-modal word representation grounded in visual context. *arXiv preprint arXiv:1711.03483*.

## 9 Appendix

### A Fine-Grained Ablation Study

In this section, we provide a more detailed ablation study based on the SimLex999 dataset for both

FastText and GloVe. Shown in Table 10, the results reveal interesting findings. The binary discrimination task ( $\mathcal{L}_B$ ) is the most beneficial one for adjectives in the case of both embeddings. This improvement arguably comes from the missing information in textual representations such as shapes, colors, and sizes of the objects which are fused by this cross-modality alignment.  $\mathcal{L}_B$  also boosts the performance of the ‘Hard’ section in which similarity is hard to distinguish from relatedness. The reason probably lies in the shift of focus toward similarity (see Table 5) which makes it easier to distinguish between similarity and relatedness. The language model tasks ( $\mathcal{L}_{FW}$  and  $\mathcal{L}_{BW}$ ) seem to contribute the most to nouns and verbs describing the scenes in the images. Moreover, our best model ( $\mathcal{L}_{All} + \mathcal{R}(\alpha, \beta)$ ), regarding all the datasets, does not achieve the best result here because each dataset focuses on a different aspect of the language (e.g, similarity or relatedness). However, our final embeddings incorporate the information from different perspectives and improve on all the datasets.

### B Refining the Textual Vector Space

Similar to the visually grounded GloVe embeddings, the grounded FastText (VGE\_F) also refine the irregularities of textual vector space (referring to Section 6). Examples of differing nearest neighbors are reported in Table 9. Since FastText performs quite well on word-level tasks, the difference is very subtle. The improvement seems to mainly fall into alleviating the antonym problem (e.g, for ‘democracy’ in the table) and clustering typos together (e.g, ‘medicine’ and ‘medecine’). We can also observe tokens such as ‘round.And’ that FastText’s tokenizer has failed to split but have been cluster together by our approach. Overall, the table confirms the results in Table 1.

democracy		possible		excited		round		medicine		flawlessly		arrogantly	
F	V	F	V	F	V	F	V	F	V	F	V	F	V
dictatorship	democracy.	necessary	possible	excited	EXCITED	round.And	round.It	medical	medecine	flawless	Flawlessly	foolishly	haughtily
		impossible	possible.So	anxious	excited-	rounded	round.The	pharmacology	pharmaceuticals				
						oval	-round	medication	medicine				
						roundin	round.Now						

Table 9: Results of 10 nearest neighbors for FastText(F) and VGE\_F (V). Only the differing neighbors are reported.

Model	SimLex999	Adjs	Nouns	Verbs	Conc-q1	Conc-q2	Conc-q3	Conc-q4	Hard
GloVe	40.8	62.2	42.8	19.6	43.3	41.6	42.3	40.2	27.2
$\mathcal{L}_B$	42.5	70.1	41.3	25.1	45.8	45.9	43.8	46.6	28.1
$\mathcal{L}_{FW}$	<b>52.6</b>	70.1	53.1	37.8	54.4	54	<b>49.3</b>	55.2	38
$\mathcal{L}_{FW} + \mathcal{L}_{BW}$	52.5	69.7	52.6	<b>40.6</b>	<b>55.5</b>	54.1	48.7	55.4	38.3
$\mathcal{L}_{FW} + \mathcal{L}_{BW} + \mathcal{L}_B$	52.5	69.8	<b>53.5</b>	37.7	53.3	53.8	48.7	<b>58</b>	<b>39.3</b>
$\mathcal{L}_{All} + \mathcal{R}(\alpha, \beta)$	51.8	<b>72.1</b>	52.0	35	53.1	<b>54.8</b>	47.4	56.8	38.3
Fasttext	47.1	59.8	50.5	31.5	46.4	<b>46.8</b>	48.5	52	29.6
$\mathcal{L}_B$	38.5	<b>64.9</b>	41.7	23.8	37.2	37	41.2	48	26.2
$\mathcal{L}_{FW}$	50.2	59	55.8	<b>37.1</b>	47.1	46.1	51.9	60.2	32.1
$\mathcal{L}_{FW} + \mathcal{L}_{BW}$	<b>50.8</b>	59.3	<b>57.3</b>	36	46.1	46.2	<b>53.1</b>	<b>62.3</b>	32.7
$\mathcal{L}_{FW} + \mathcal{L}_{BW} + \mathcal{L}_B$	<b>50.8</b>	60.6	57.1	35.8	<b>48.1</b>	46.4	52.7	61.3	<b>33.4</b>
$\mathcal{L}_{All} + \mathcal{R}(\alpha, \beta)$	49.0	58.6	54.1	32.9	45.3	46.7	51.3	57.7	31.3

Table 10: Fine-grained ablation study on SimLex999 (Spearman’s  $\rho$ ). Conc-q1 and Conc-q4 contain the most abstract and concrete words respectively. The hard section includes a set of word-pairs in which similarity is hard to distinguish from relatedness