**ORIGINAL MANUSCRIPT**

# Language with vision: A study on grounded word and sentence embeddings

Hassan Shahmohammadi[1] · Maria Heitmeier[1] · Elnaz Shafaei-Bajestan[1] · Hendrik P. A. Lensch[1] ·
R. Harald Baayen[1]

## Abstract

Grounding language in vision is an active field of research seeking to construct cognitively plausible word and sentence representations by incorporating perceptual knowledge from vision into text-based representations. Despite many attempts at language grounding, achieving an optimal equilibrium between textual representations of the language and our embodied experiences remains an open field. Some common concerns are the following. Is visual grounding advantageous for abstract words, or is its effectiveness restricted to concrete words? What is the optimal way of bridging the gap between text and vision? To what extent is perceptual knowledge from images advantageous for acquiring high-quality embeddings? Leveraging the current advances in machine learning and natural language processing, the present study addresses these questions by proposing a simple yet very effective computational grounding model for pre-trained word embeddings. Our model effectively balances the interplay between language and vision by aligning textual embeddings with visual information while simultaneously preserving the distributional statistics that characterize word usage in text corpora. By applying a learned alignment, we are able to indirectly ground unseen words including abstract words. A series of evaluations on a range of behavioral datasets shows that visual grounding is beneficial not only for concrete words but also for abstract words, lending support to the indirect theory of abstract concepts. Moreover, our approach offers advantages for contextualized embeddings, such as those generated by BERT (Devlin et al, 2018), but only when trained on corpora of modest, cognitively plausible sizes. Code and grounded embeddings for English are available at (https://github.com/Hazel1994/Visually_Grounded_Word_Embeddings_2).

**Keywords** Visual grounding · Multi-modal word embeddings · Grounded cognition · Grounding abstract words

## Introduction

Where do symbolic representations of language get their meaning from? It has been argued both from a theoretical and an empirical perspective that knowledge is grounded in perceptual experience (Barsalou, 2008; Lakoff, 1987; Langacker, 1999; Zwaan and Madden, 2005). Evidence for this embodied view of knowledge comes from a range of scientific domains such as neuroimaging (e.g., Simmons et al, 2005; Martin, 2007) and behavioral studies (e.g., Goldstone, 1995; Solomon and Barsalou, 2001; Solomon and Barsalou, 2004), showing that knowledge is grounded in sensory, but also interoceptive perception and motor action (overview in Barsalou, 2008). Nevertheless, this perspective is not without opposition. For example, Louwerse and Connell (2011) argue that linguistic information suffices for more shallow processing of meaning and that perceptual, embodied information is only accessed when deeper knowledge of a word is required.

This debate has been stimulated further by the success of meaning representations which are based on linguistic information alone. They build on the notion of Harris (1954) that similar words occur in similar contexts and represent each word as numerical vectors, with similarities between these vectors reflecting similarities in words' meanings. By now, many different methods have been devised to generate such vectors (called "word embeddings" in natural language processing (NLP) and throughout the remainder of this paper), beginning with Hyperspace Analogue to Language (HAL;

✉ Hassan Shahmohammadi
   hassan.shahmohammadi@uni-tuebingen.de

1  University of Tübingen, Tübingen, Germany

Lund and Burgess, 1996) and latent semantic analysis (LSA; Landauer and Dumais, 1997), and later, mainly in the fields of NLP and machine learning, Word2Vec (Mikolov et al, 2013), Fasttext (Bojanowski et al, 2017) or GloVe (Pennington et al, 2014). Today, word embeddings are employed successfully in many different areas and tasks within NLP, such as POS tagging, named-entity recognition, and sentiment analysis (Wang et al, 2019).

As an easily obtained representation of semantics, word embeddings are also used in many areas of cognitive science, such as AI research, psychology, or psycholinguistics, with encouraging results (see Günther et al, 2019). From a cognitive perspective, word embeddings have been evaluated in two ways. A relatively direct method is to compare them to metrics obtained from brain imaging such as fMRI or EEG. Bulat et al (2017); Hollenstein et al (2019) showed that a variety of word embeddings (e.g., GloVe, Word2Vec, Fasttext) correlate relatively well with such metrics. A second, more indirect, approach uses behavioral data such as reaction times or ratings as evaluation criteria. Mandera et al (2017) showed that word embeddings can be used to predict semantic priming as well as word associations, similarity/relatedness ratings and even perform well in a multiple-choice task. Further evidence in favor of the cognitive plausibility of word embeddings has been provided by Westbury (2014); Westbury and Hollis (2019) who predicted familiarity and humor ratings, respectively, Marelli and Amenta (2018) who demonstrated that the semantic relatedness of words' orthographic neighbors is predictive for visual lexical decision and naming latencies, Abdou et al (2021) who showed that even color relations are accurately represented by purely textual embeddings, as well as Louwerse and Zwaan (2009); Avery et al (2021); Gatti et al (2022) who demonstrated that geographical locations of cities are reflected in purely textual embeddings. Recently, embeddings have also found their way into psycholinguistic models. For example, the Discriminative Lexicon Model (Baayen et al, 2019; Heitmeier et al, 2021, 2023), a model of the mental lexicon, uses word embeddings to represent words' meanings. Other models also use distributional information to represent semantics, either randomly generated ones (e.g., Gaskell and Marslen-Wilson, 1997; Magnuson et al, 2020) or based on human ratings (e.g., mir, 2008), further highlighting the need for a large set of psychologically valid word embeddings. However, the cognitive plausibility of mechanisms generating word embeddings such as *Word2Vec* has not gone unchallenged (Mannering and Jones, 2021).

While the success of textual embeddings has nevertheless led some researchers to believe that meaning can be fully, or at least to a large extent, be derived from language alone (Landauer, 1999), the wide range of empirical evidence in favor of a grounded view of knowledge representation and cognition has sparked the search for representations that are informed not only by text, but also by vision and other modalities (see also Andrews et al, 2014).

Therefore, a number of previous studies have tried to improve word embeddings by using available data similar to text corpora. Some studies have tried to extract meaning representations exclusively from visual information (usually images). The resulting visual word embeddings have been found to be very good models of human perceptual behavior (e.g., Zhang et al, 2018), but success at predicting other behavioral data was more mixed, with some reporting positive (Lüddecke et al, 2019; Bulat et al, 2017) and others negative results compared to textual embeddings (e.g., Peterson et al, 2017; De Deyne et al, 2021; Rotaru and Vigliocco, 2020; Utsumi, 2022). The more promising approach has been to ground textual embeddings in vision, i.e., to include visual information with textual embeddings. The resulting embeddings are usually referred to as multimodal embeddings. This approach is especially promising because textual and visual representations seem to carry different kinds of information (Petilli et al, 2021; Andrews et al, 2014). Multimodal embeddings have been successful in a range of areas. They have been shown to correlate better than purely textual embeddings with human similarity/relatedness judgments and concept categorization. Bulat et al (2017); Anderson et al (2015) found that they are better at predicting brain activity than purely textual embeddings. Moreover, they are useful in modeling the learning of novel words' meanings in both children and adults (Lazaridou et al, 2016, 2017). Finally, they have been shown to improve performance in a number of classification tasks in NLP (Bordes et al, 2019).

Several approaches to obtaining multimodal embeddings are available. We restrict our discussion here to approaches combining textual and visual information, but a body of work has also explored the integration of emotional (e.g., Rotaru and Vigliocco, 2020), sensory (e.g., Johns and Jones, 2012), auditory (Kiela and Clark, 2015) and olfactory (Kiela et al, 2015) information. Early approaches gleaned visual information from human ratings, e.g., by utilizing data collected in the ESPGame dataset (Von Ahn, 2006), or used "Bag-of-Visual-Word" approaches where images are chunked into small pieces to form a kind of visual vocabulary (e.g., in Anderson et al, 2015). More recently, feature vectors have been extracted directly from computer vision models (see Baroni, 2016 for a review).

Subsequently, the visual information needs to be combined with textual information. Baroni (2016) differentiates between two approaches: *cross-modal mapping* and *multimodal fusion*. Cross-modal mapping approaches to grounding textual in visual information aim to map between one and the other, in an attempt to account for how vision could be translated into language or vice versa (Baroni, 2016). An early model inferring perceptual embeddings by linking words using distributional semantics is Johns and Jones

(2012). They used feature norms from McRae et al (2005) to model perceptual representations. For words for which no feature norms were available, they inferred these by first computing the similarity of the target word with all words for which feature norms were available using distributional semantics, and then computed a weighted average of their feature norms. After having inferred feature norms for all words, they repeated the process in a second step, this time taking into account all words, rather than only those for which feature norms were available originally. A more recent proposal for connecting textual and visual embeddings by means of a simple linear mapping can be found in Günther et al (2022).

On the other hand, multimodal fusion (Baroni, 2016) aims to combine textual and visual information into a single representation. The simplest example for multimodal fusion is concatenation, as is often used when multimodal embeddings are explored in cognitive science and psychology (e.g., Utsumi, 2022; Rotaru and Vigliocco, 2020). However, there are also more sophisticated approaches from the realm of NLP: Some approaches apply feature-level fusion, combining image features with textual word embeddings (after obtaining both separately) with methods such as singular value decomposition (SVD) or gated recurrent units (GRU) (Cho et al, 2014b; Bruni et al, 2014; Kiela and Bottou, 2014; Kiros et al, 2018). Others learn multimodal word representations in a joint feature space defined by a specific criterion (known as loss function) between modalities, for example by using auto-encoders (Silberer and Lapata, 2014; Hasegawa et al, 2017) or long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) networks (Kiela et al, 2018; Chrupała et al, 2015). Recently, new approaches based on modality alignment have emerged. Here, vision and language are treated separately (as opposed to having both in a shared space) but the textual embeddings are aligned with image features (Shahmohammadi et al, 2021; Bordes et al, 2019).

In the present work, we make use of recent advances in machine learning, computer vision and NLP to propose a new method of computing multimodal embeddings via multimodal fusion (Baroni, 2016). Our approach falls into the latter category of grounding models where rather than projecting textual and visual embeddings into the same space, textual embeddings are slightly adjusted to reflect information gleaned from images (see Fig. 1). Our model is able to generalize to new words without a visual representation, which allows it to generate grounded embeddings not only for concrete words for which images are available but also for abstract words, extending earlier work such as Johns and Jones (2012); Utsumi (2022) while making use of more recent insights from NLP. We compare our model to both ungrounded embeddings as well as embeddings based on other grounding methods and show that our model is more predictive for responses in a range of behavioral datasets

such as similarity/relatedness judgements (e.g., MEN, Bruni et al, 2014) which have been used in previous work to evaluate word embeddings from a psycholinguistic perspective (Mandera et al, 2017). Our grounded embeddings are made available to the community.[1]

Our grounded embeddings allow us to explore various questions which arise from previous work on grounding and generating distributed meaning representations in general, and which are crucial when aiming to model cognitively plausible meaning representations:

1. On the one hand, many studies have shown that combining visual information and textual information is attractive from a theoretical point of view (e.g., Andrews et al, 2014; Lake and Murphy, 2021) and indeed improves the quality of word embeddings (e.g., Bruni et al, 2014; 2021, 2021; Lazaridou et al, 2016). On the other hand, purely textual embeddings are very successful even on tasks related to vision and spatial relations (Louwerse and Zwaan, 2009; Abdou et al, 2021), and purely visual embeddings do not perform well at predicting human similarity judgments (e.g., De Deyne et al, 2021). Hence, the extent to which textual representations benefit from visual grounding, as well as the specific tasks and methods that are most effective, remains an open question. Apparently, a fine balance has to be struck between too much and too little visual information in grounding. A number of studies has attempted to explore this question from both a more technical, engineering perspective, but also from a cognitively motivated perspective. For instance, Hill and Korhonen (2014); Rotaru and Vigliocco (2020) found that how beneficial perceptual information is for resulting embeddings depends on the concreteness of the words: the more concrete the words are, the more they profit from perceptual information. We will explore to what extent perceptual knowledge from images is beneficial for acquiring high-quality and cognitively plausible embeddings, using a more modern grounding architecture.

2. Traditionally, embeddings are grounded on a single word basis (e.g., Günther et al, 2022; Kiela and Bottou, 2014; Bruni et al, 2014). However, visual scenes are complex, and are usually best described not by single words, but rather by entire sentences. Equating complex scene structures with isolated words is not only counter-intuitive but also problematic when grounding abstract words since highly abstract words (e.g., *justice*) are rarely depictable. It is known that language is vital for representing abstract concepts (Borghi et al, 2017; Dove, 2018). However, the interplay between language and perceptual experiences is

---

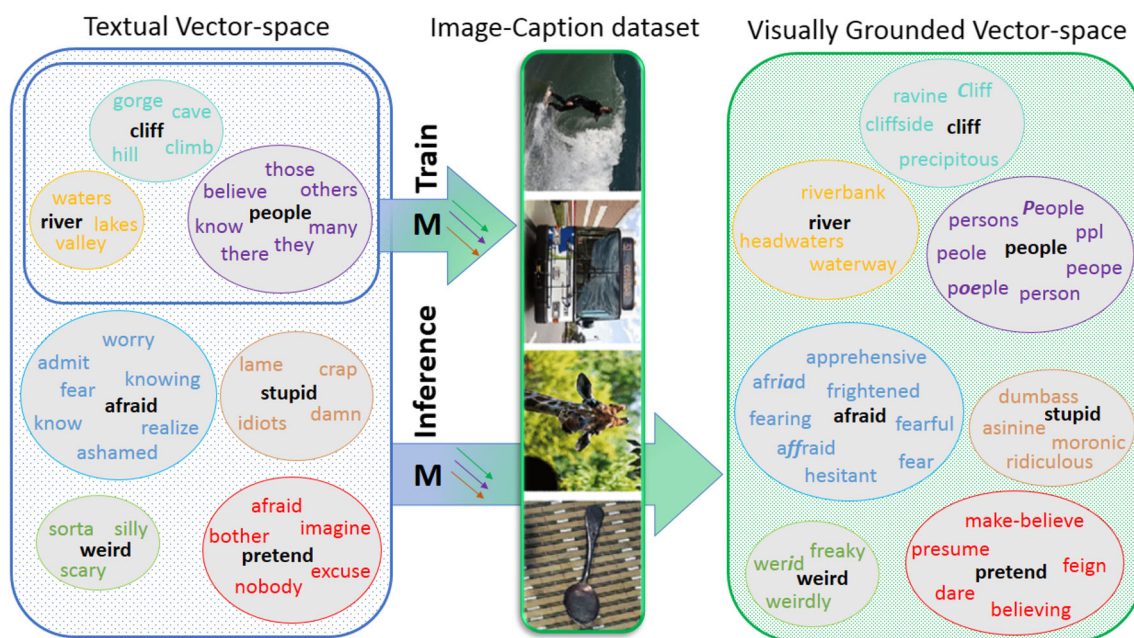[1] https://github.com/Hazel1994/Visually_Grounded_Word_Embeddings_2

**Fig. 1** Our model constructs visually grounded embeddings (*right*) from textual embeddings (*left*) by applying a learned alignment (*M*) trained on a subset of 10,000 words in image–caption pairs. It then generates zero-shot grounded embeddings at the inference phase for a total of 2,000,000 words, including not only concrete words but also abstract words. For each query word (in *black*), the grounded embed-dings (*right*) retrieve more similar words compared to the purely textual embeddings (*left*) and alleviate the bias toward dissimilar words with high co-occurrence frequencies such as (*many, people*). Out of the top ten nearest neighbors for each query word, only the differing neighbors between the textual embeddings and the grounded embeddings are shown in the right-hand panel

still an open field. How do language and embodied experience together shape our understanding of abstract and concrete concepts? We will design various experiments to explore how language (here represented as word representations) and vision (images) should interact.

3. There exist multiple theories of how words are grounded in perceptual experiences (Paivio, 1971; Borghi et al, 2019; Howell et al, 2005). Nonetheless, large-scale grounding of abstract words into vision is still an open field. More specifically, the question still remains: how should abstract words be grounded in computational models on a large scale? In line with the theory of indirect grounding (Howell et al, 2005; Louwerse, 2011), we propose a large-scale grounding method[2] to effectively ground abstract words.

4. Newly proposed large-scale contextualized language models rely on enormous amounts of data (e.g., BERT: Devlin et al, 2018). While this leads to good performance, it is cognitively implausible, as humans encounter only a much smaller number of words over their lifetimes (Brysbaert et al, 2016). Our fourth question therefore relates to whether visual grounding is equally helpful when large amounts, or only small amounts, of training

data are available: How much does the amount of training data influence the improvement of visual grounding on downstream tasks such as sentiment analysis? We will demonstrate that on corpora sizes closer to human-scale training data, visual grounding improves the quality of embeddings even on highly abstract tasks.

To this end, our paper is structured as follows. Sections 2 and 3 introduce our method, which is evaluated in Section 4. In Sections 5 and 6 we will address the first two aforementioned research questions. Furthermore, we will investigate the impact of grounding on task performance, specifically in state-of-the-art language processing models, with respect to the available training data in Sections 7 and 8.

## Visually grounded word embeddings

In this section, we explain our visual grounding approach and how it can be used to generate visually grounded word representations from textual word embeddings. For $(S_j, I_j) \in D$, let $S_j = [w_1, w_2 \cdots w_n]$ be a textual caption with $n$ words describing its corresponding image with the image vector $I_j$ in the dataset $D$. The image vector $I_j$ is obtained by feeding the image into a pre-trained convolutional neural net-

---

[2] Please note that our model is not a cognitive model. However, our findings provide substantial support for the indirect grounding theory.

work (CNN) model. A CNN is a family of neural networks designed for processing images with a grid-like topology that extracts local information and aggregates them through multiple layers of learnable parameters. CNNs are usually trained on a large set of images annotated by human raters to classify images into many classes (e.g., *dog*, *horse*, and *car*). Once they are trained, they can be used to encode images into dense and meaningful numerical representations that correspond well to human intuitions (Bracci et al, 2019; Lazaridou et al, 2017). Let $t_i \in \mathbb{R}^d$ be a textual embedding of the word $w_i$, which has been obtained by a pre-trained word embedding model $T_e : w_i \mapsto t_i$ (e.g., Fasttext). The goal is to learn a linear mapping $M$ to visually ground any textual word vector $t_i$ in its corresponding image vector $I_j$ and obtain the visually grounded embedding $g_i \in \mathbb{R}^c$ of the word $w_i$. The learned mapping $M$ will linearly adjust the textual word embeddings based on the information in images. This mapping ideally should: a) preserve the abstract knowledge from co-occurrence statistics captured by textual embeddings trained on large textual corpora, and b) align the textual embeddings with their corresponding visual properties available in images. This way, the grounded embeddings will benefit both concrete and abstract words (Shahmohammadi et al, 2021). While it may seem intuitive to learn both modalities in a shared feature space, we argue that such approaches, unfortunately, are more likely to cause the grounded embeddings to lose the abstract knowledge from textual co-occurrences and therefore suffer from a bias towards concrete words as reported by Park and Myaeng (2017).

It is widely acknowledged that language plays a crucial role in acquiring abstract concepts (Borghi et al, 2017; Dove, 2018). Therefore, we believe that preserving abstract knowledge during the grounding process requires individual words to be aware of the context (other words in the sentence). The grounding process should also respect the textual vector space as any random change to textual embeddings will distort the semantic information obtained by textual statistics (Shahmohammadi et al, 2021). Figure 2 lays out the architecture of our proposed grounding model. The grounded version of any word $w_i$ is obtained by mapping its textual embedding $t_i$ into the visually grounded space using the linear mapping $M$ as $g_i = t_i \cdot M$. In the grounded space word vectors are aligned with the images by using a one-layer long short-term memory (LSTM) network (Hochreiter and Schmidhuber, 1997). The $LSTM$ network is a type of recurrent neural network that is suitable for text processing, it processes a sequence of words in a text word by word and at each step updates its internal learning parameters. The LSTM encodes the whole sentence $S_j$ as a single vector $h_n$:

$$h_n = LSTM(G, c_0, h_0 \mid \theta), \qquad (1)$$

where $G$ denotes the input — all the grounded word vectors (output of $M$) — and $\theta$ the learning parameters. It also includes a cell state $c_t$ and a hidden state $h_t$ where $t$ denotes the current time-step (the current word being processed). At first, the network is initialized with a random hidden and cell states ($h_0$ and $c_0$) and takes one word at each time-step (see
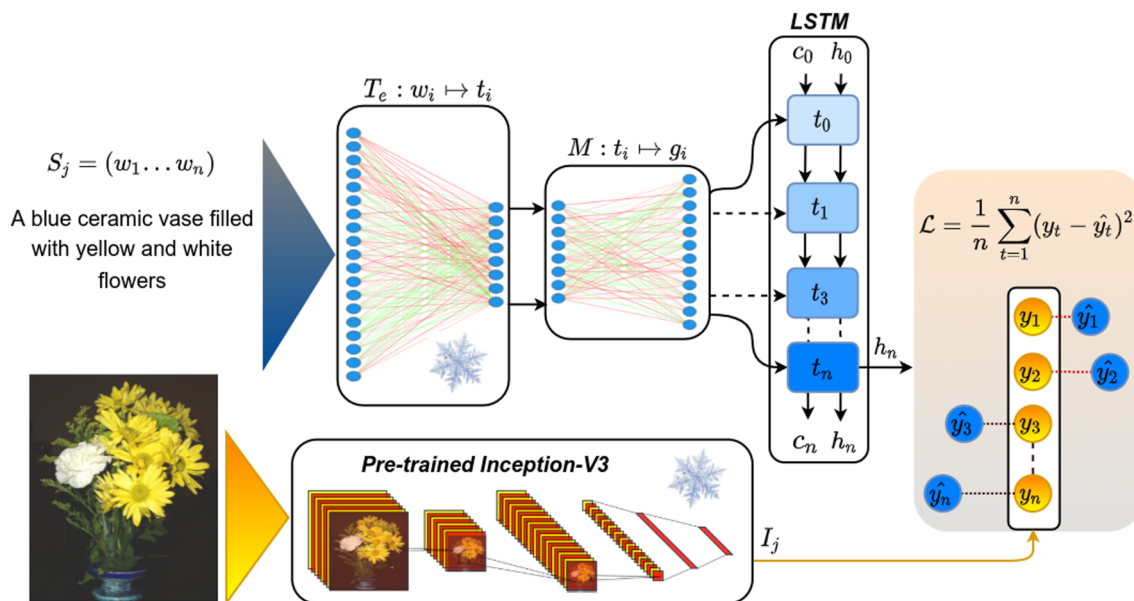


**Fig. 2** Our visual grounding model encodes each caption word by word, using an LSTM, given the task to predict the corresponding image vector. A mapping M is set up that takes textual vectors and maps them into the grounded space. This mapping is trained on a limited number of words, those that occur in the captions, but is then applied to all the words, after the training is completed, to generate "zero-shot" (unseen) grounded embeddings. The *snowflake icon* indicates the frozen learning parameters during training

Fig. 2) and each time, for each successive word, it updates its memory by removing and adding information to the cell state. It then generates an output $h_t$ based on the current input $g_t$ and $c_t$. Both $h_t$ and $c_t$ are passed to the next time-step. We extract the output of the last time-step $h_n$ as a vector representing the whole sentence. The model is trained to match $h_n$ to the image vector $I_j$ for each particular training sample $(S_j, I_j) \in D$. We optimize the parameters of the $LSTM$ and the mapping $M$ (denoted as $\Theta$) based on the following mean-squared-error (MSE) loss:

$$\hat{\Theta} = \underset{\Theta}{\mathrm{argmin}} \frac{1}{N} \sum_{t=1}^{n} (y_t - \hat{y}_t)^2, \qquad (2)$$

where $y$ and $\hat{y}$ denote the ground truth image vector ($I_j$) and the predicted image vector ($h_n$), respectively. By applying the $LSTM$ network, the model takes into account the context in which each word occurs. Therefore, the whole sentence is mapped to the image vector. Since the model tries to predict an image vector, it will change the textual vector space such that the image vector is estimated as accurately as possible. Nonetheless, we restrict the influence of the images on the word vectors by keeping the mapping $M$ linear. Naturally, the grounded word vectors (output of $M$) will still respect the textual vector space but they will be indirectly aligned to the image representations.

After training the model on (caption, image) pairs, the mapping $M$ can be used to indirectly ground both abstract and concrete words including out-of-vocabulary words. For instance, for obtaining the visually grounded vector of the word *sad*, we first fetch its textual vector $t_{sad}$ using the pre-trained textual embeddings. The grounded vector is then obtained by using the learned mapping $M$ as $g_{sad} = t_{sad} \cdot M$, where $g_{sad}$ indicates the visually grounded version of the word *sad*. In this way, a visually grounded version of the textual embeddings is created in a zero-shot manner (including unseen words) despite being exposed to only a limited number of words while training on image captions.

## Implementation details

We used the Microsoft COCO 2017 dataset (Lin et al, 2014) in our experiments. Each sample of this dataset includes a single image along with five different human-generated captions (Chen et al, 2015). The whole dataset was divided into 118,000 train and 5000 validation samples. We set the batch size to 256 with each batch containing 256 image vectors (of dimension 2048) along with one of their corresponding captions. Image vectors were extracted from the penultimate layer of a pre-trained Inception-V3 CNN model (Szegedy et al, 2016), based on ImageNet (Deng et al, 2009). We set

the dimension of the grounded embeddings (output of $M$) to 1024, following Shahmohammadi et al (2021). A one-layer LSTM was applied with 2048 units. We removed the punctuation marks from the captions and converted all words to lowercase. Only the top 10,000 most frequent words in the captions were used and the rest were ignored. Reducing the number of processed words is a common practice in NLP, as many words occur rarely in the training corpus and therefore make a negligible contribution to the learning process. We trained the model for 20 epochs (20 iterations on the whole dataset) with five epochs tolerance early stopping, using the NAdam optimizer (Dozat, 2016) with a learning rate of 0.001. Early stopping is a technique to prevent a model from overfitting to the training data by stopping the training process once the model's performance on a validation dataset stops improving. In our setup, we train the model until its validation score decreases for five consecutive epochs, after which the training process is halted using early stopping.

Both the pre-trained textual embedding $T_e$ and the Inception-V3 model are frozen—weights are kept fixed—during training. Two popular pre-trained textual word embeddings, GloVe ($crawl - 300d - 2.2M - cased$) and Fasttext ($crawl - 300d - 2M - SubW$), were used to initialize the embedding $T_e$. Therefore, we generated two sets of grounded embeddings, one from Fasttext and one from GloVe.

## Evaluation

In this section, we develop several evaluation techniques to study the behavior of visually grounded embeddings and address the initial question of how much and in what specific applications perceptual information from images contributes to the creation of high-quality and cognitively plausible embeddings.

### General evaluation

The question of how to appropriately evaluate word embeddings persists, despite the existence of numerous evaluation benchmarks (Wang et al, 2019). However, in both psycholinguistics and NLP, humanly annotated lexical semantic similarity or relatedness datasets are commonly used to evaluate (multi-modal) embeddings (Mandera et al, 2017; Rotaru and Vigliocco, 2020; De Deyne et al, 2021; Park and Myaeng, 2017). Here, the task is to estimate the similarity/relatedness score of a given pair of words with the Spearman correlation as evaluation metric. Relatedness is based on topical match which quantifies the degree to which two words are associated with each other (*child-play*). Similarity is based on taxonomic closeness which is a subset of relatedness and quantifies how alike two words are (*car-automobile*). It is worth noting that some datasets do not distinguish between

similarity and relatedness. For example, the pair (*clothes*, *closet*) comes with the score of 1.96 (out of 10) in Sim-Lex999, but exactly the same pair receives a score of 8.00 in WordSim353, which does not distinguish between similarity and relatedness. We assess the quality of our visually grounded word representations using the following datasets and juxtapose the results with textual embeddings and related previous works.

**MEN** (Bruni et al, 2014) This dataset is compiled specifically for the purpose of evaluating multi-modal models. It only contains words that appear as image labels in the ESP-Game[3] and MIRFLICKR-1M16[4] datasets. Therefore, it is suitable for multi-modal assessments. MEN consists of 3000 word pairs with semantic relatedness ratings obtained via Amazon Mechanical Turk. For example, (*sun*, *sunlight*) has a MEN score of 50 (out of 50) but the score of (*zebra*, *bakery*) is 0.

**WordSim353** (Finkelstein et al, 2001) This collection contains 353 word pairs annotated by 13–16 human judgments for each pair. The judges did not distinguish between similarity and relatedness. For instance, (*computer*, *keyboard*) comes with a score of 7.62 (out of 10).

**SimLex999** (Hill et al, 2015) Unlike WordSim353, Sim-Lex999 draws a clear distinction between similarity and relatedness as mentioned above. SimLex999 contains 999 word pairs annotated by 500 annotators via Amazon Mechanical Turk. Both WordSim353 and SimLex999 have been used for explaining human performance in psycholinguistic tasks (Mandera et al, 2017).

**Rare-Words** (RW, Luong et al, 2013) This dataset measures the performance of a word-embedding model on rare words that occur less frequently (based on Wikipedia). It contains 2034 word pairs annotated by ten human judges. Examples of words in this collection are *interjection* and *behaviorist*.

**MTurk771** (Halawi et al, 2012) MTurk771 consists of 771 word pairs. The authors used WordNet[5] to extract both related and unrelated word pairs and collected 20 human ratings for each word pair.

**SimVerb3500** (Gerz et al, 2016) This dataset provides human ratings for the similarity of 3500 *verb* pairs. Providing broad coverage of verbs, this dataset offers a great resource for a better understanding of "the complex diversity of syntactic-semantic verb behaviors" (Gerz et al, 2016, p. 2174).

---

[3] http://www.cs.cmu.edu/~biglou/resources/

[4] https://press.liacs.nl/mirflickr/

[5] https://wordnet.princeton.edu/

Table 1 shows the evaluation results on lexical semantic benchmarks. Our zero-shot grounded embeddings are shown as ZSG-G and ZSG-F indicating the grounded versions of GloVe and Fasttext, respectively. The initial segment of the table demonstrates that ZSG-G exhibits superior efficacy compared to textual GloVe across *all* benchmarks. In the case of Fasttext on the other hand, improvements are somewhat more modest, probably because Fasttext takes into account sub-word information. That is, it takes advantage of the internal structure of a word to improve vector representations. For instance, the word vector of *eating* might be a combination of the *eat* and *ing*. Hence, it might capture word similarity/relatedness better compared to GloVe which treats each word as a unique item. In the lower part of the table, we compare the performance of our best model (ZSG-G) with related visually grounded embedding models. For a fair comparison, we limit our list to those who adopted pre-trained word embeddings. Shahmohammadi et al (2021) (shown as VGE-G in the table) proposed a similar grounding approach to ours where they train a linear mapping to transfer from textual word representations to visually grounded representations. However, the main difference with our approach is the training scheme of the mapping. While we only train using a single task (predicting the associated image vector given its caption), multitask training with three different tasks is adopted in their approach. In their setup, the model generates the corresponding caption word by word for a given image vector in both forward and backward directions. Furthermore, the model receives pairs of captions and images as inputs and learns to discriminate between matching and non-matching pairs. While inspired by their method, our approach is simpler, requires less computational power, and performs slightly better on the same set of benchmarks.

Kiela et al (2018) also proposed a visual grounding approach for pre-trained textual word representations (GloVe), by using the same image database as ours. Similar to Shahmohammadi et al (2021) their approach is based on multitask training where the following tasks have been proposed: Cap2Img: predicting the image vector from its caption; Cap2Cap: generating an alternative caption of the same image; Cap2Both: training by Cap2Cap and Cap2Img simultaneously. Our approach, despite its simplicity, captures the semantic relationships of words much better compared to Cap2Both and Cap2Img. Next, we compared our results with polymodal embeddings by Park and Myaeng (2017). In this approach, the meaning of each word is derived from six different types of distinct embeddings including linear context, syntactic context, visual perception, cognition, emotion, and sentiments based on the human cognitive model proposed by Maruish and Moses (2013). Even though their approach uses more resources including two pre-trained embeddings (Word2Vec, GloVe) and incorporating other modalities, ours is still superior on MEN and WSim353, albeit worse on

**Table 1** Comparison of our grounded embeddings (ZSG-*) to textual embeddings and other visually grounded embedding models

| 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 |
| ZSG-G (ours) | **53.2** | ***85.1** | ***78.8** | ***73.2** | ***38.5** | ***52.6** | **63.6** |
| Fasttext | 56.1 | 81.5 | 72.2 | ***75.1** | 37.8 | 47.1 | 61.6 |
| ZSG-F (ours) | ***57** | ***84.4** | **72.3** | 74.5 | ***39.6** | ***49.6** | **62.9** |
| VGE-G | 52.6 | **85.1** | **78.9** | ***73.4** | 37.4 | 51.8 | 63.2 |
| ZSG-G (ours) | **53.2** | 85.1 | 78.8 | 73.2 | ***38.5** | ***52.6** | **63.6** |
| Cap2Both | 48.7 | 81.9 | 71.2 | _ | _ | 46.7 | |
| Cap2Img | 52.3 | 84.5 | 75.3 | _ | _ | 51.5 | |
| Park & Myaeng | _ | 83.8 | 77.5 | _ | _ | **58.0** | |
| P&M_VG. | _ | _ | _ | _ | _ | 15.7 | |
| Collell et al. | _ | 81.3 | _ | _ | 28.6 | 41.0 | |

Our embeddings show stronger correlation with human ratings on most of the datasets. The metric is Spearman's $\rho \times 100$. Number with stars indicate statistically significant differences ($p < 0.05$ *; $p < 0.01$ **; $p < 0.001$ ***, $t$ tests) between our grounded embeddings (ZSG-G) and textual (GloVe or Fasttext) or VGE-G embeddings

Simlex999. The large performance gap observed for Sim-Lex999 may be attributed to the multi-modality training of the model conducted by Park and Myaeng (2017). Employing solely their visually grounded embeddings (P&M_VG) results in low-quality word vectors, further confirming that their visually grounded embeddings do not benefit abstract words (Park and Myaeng, 2017).

For further consolidation, we calculated the $t$ test[6] (Student, 1908) between the predictions of textual and grounded embeddings for both GloVe and FastText and compared the results of our grounded GloVe (ZSG-G) with the previous VGE-G by Shahmohammadi et al (2021) (denoted as *, **, or *** in Table 1). All the improvements over the textual embeddings were found to be statistically significant with the exception of *RW* dataset using GloVe. The differences in performance between our embeddings and VGE-G were found to be significant across all the benchmarks.

In summary, our approach, while trained on a limited number of words available in image captions, creates visually informed word representations, even for unseen words, which are more aligned with human judgment across a wide range of human-rated word similarity and relatedness tasks.

In linguistics, concrete words[7] refer to physically real and perceptible entities such as *tree*, *ball*, or *Chris*, whereas abstract words have references that are not readily perceptible to the senses, and are more complex and variable in mean-

ing, including mental states (e.g., *happiness*), events (e.g., *encounter*), conditions (e.g., *totalitarianism*), relations (e.g., *brotherhood*) and so forth (VandenBos, 2015; Borghi and Binkofski, 2014; Barsalou et al, 2018; Davis et al, 2020). Concreteness and abstractness are not binary properties of words (Wiemer-Hastings et al, 2001). Words become increasingly abstract as they are more separated from physical entities and more linked to mental states (Barsalou, 2003). Word concreteness indicates the degree to which a word denotes a perceptible entity and is measured on a numerical scale by subject ratings (Brysbaert et al, 2014). For example, the word *pancake* is ranked high on the scale as it is associated with many sensory properties such as smell, taste, shape, color, etc.

Extensive evidence from behavioral experiments suggests that there is an advantage in cognitive processing of words for concrete over abstract words—often referred to as the "concreteness effect". It has been shown that concrete words, compared to abstract words, are processed faster in isolation (Schwanenflugel and Shoben, 1983) and non-supportive contexts (Schwanenflugel and Stowe, 1989), are remembered better in paired associative learning (Paivio, 1965) and free recall tasks (Schwanenflugel et al, 1992), and are learned faster (Mestres-Missé et al, 2014). Evidence has been put forward for this distinction in the brain. Case reports of patients with brain damage demonstrate differential impairments with regard to abstract and concrete concepts (Breedin et al, 1994; Tyler et al, 1995; Warrington, 1975). Neuroimaging studies provide evidence for overlapping but distinct brain areas engaged in the processing of abstract and concrete concepts (see Montefinese, 2019 for a review).

To investigate the influence of grounding on abstract and concrete words, we leverage the SimLex999 dataset. It

---

[6] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats. ttest_ind.html

[7] We assume individual words, as they are realized in English writing conventions, are the verbal expression of lexical concepts in language, and thus the terms "word" and "concept" are used interchangeably in this section.

**Table 2** SimLex999 (Spearman's $\rho \times 100$) results

| 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 | 51.8 | 72.1 | 52.0 | **35** | 53.1 | **54.8** | 47.4 | 56.8 | 38.3 |
| ZSG-G (ours) | **52.6** | **73.8** | **53.1** | 34.6 | **54.6** | 53.9 | 48.1 | 59.2 | **39.3** |
| PictureBook | 37.3 | 11.7 | 48.2 | 17.3 | 14.4 | 27.5 | 46.2 | **60.7** | 28.8 |
| PictureBook+GloVe | 45.5 | 46.2 | 52.1 | 22.8 | 36.7 | 41.7 | **50.4** | 57.3 | 32.5 |

Conc-q1 and Conc-q4 indicate the most abstract and concrete words, respectively. Our model (ZSG-G) demonstrates stronger associations with human annotators' similarity ratings on multiple categories

divides its words into different categories including adjectives, nouns, verbs, concreteness quartiles (from 1 to 4 increasing the degree of concreteness), and 'hard' sections. The 'hard' section includes the 333 most associated word pairs in the University of South Florida Free Association Database (USF) (Nelson et al, 2004). This subset of SimLex999 is reported to be the hardest for semantic models to capture because the noise from the high association makes it hard to distinguish between similarity and relatedness (Hill et al, 2015). Examples of this category are *happy-cheerful* and *weird-strange*. Table 2 shows our fine-grained evaluation on SimLex999. We compared our fine-grained results with that of Picturebook, another kind of visually grounded embeddings (Kiros et al, 2018). For each word, PictureBook retrieves the top-k images using image search. The retrieved images are then passed through a CNN trained with a semantic ranking objective with 100+ million images (Wang et al, 2014). The grounded embedding of each word is computed based on a combination of image vectors and the pre-trained GloVe embedding of that word. Our best model (ZSG-G) captures semantic relationships much better compared to other visually grounded embeddings and generalizes across different word types. For example, it not only demonstrates a more pronounced association with highly concrete (Conc-q4) words by a margin of 19.2 percentage points, but also with highly abstract words (Conc-q1) by a margin of 11.3 percentage points compared to the textual GloVe vectors. In contrast, PictureBook (Kiros et al, 2018), for example, highly benefits the more concrete words but adversely affects the more abstract category even when combined with GloVe embeddings. In comparison with VGE-G by Shahmohammadi et al (2021), our model again achieves better results while being much simpler and less computationally expensive.

We further extended the analysis of abstract and concrete words by using all the word similarity/relatedness datasets. For this aim, we first combined all the datasets (see Section 4) after normalizing the score of each dataset. That is, we transformed the scores to be in the range of [0, 1] as follows:

$$x_{in} = \frac{x_i - min}{max - min},$$

where $x_{in}$ and $x_i$ indicate the new score and the original score of the $i$th word pair, respectively. *max* and *min* denote the maximum and minimum scores within the given dataset. After normalizing and combining all the benchmarks we obtained 10,657 word pairs. We then ranked all the word pairs based on a concreteness rating dataset compiled by Brysbaert et al (2014). This dataset contains 37,000 words and 3000 two-word phrases rated by over 4000 subjects using the Amazon Mechanical Turk (MTurk) crowdsourcing platform. We denote this dataset as MTurk40k. We took the intersection between MTurk40k and our combined dataset which resulted in 8936 word pairs with both similarly/relatedness and concreteness scores. We refer to this dataset as *WCR* (word concreteness rating) for simplicity. The concreteness score of a word pair was obtained by taking the average scores of its constituent words. Examples of highly abstract and concrete word pairs from *WCR* are *(belief, purpose)* and *(apple, lemon)*, respectively. Having access to a large set of word pairs with concreteness scores, we can more thoroughly assess the behavior of visual grounding on abstract
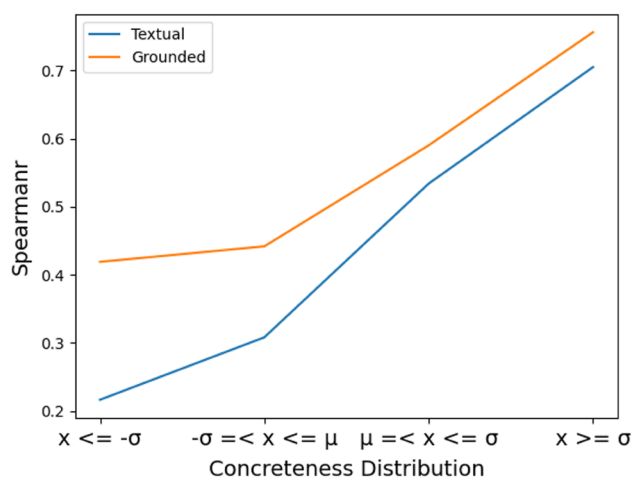


**Fig. 3** Comparison between textual and grounded embeddings of word pairs with different concreteness scores. Visually grounded embeddings highly benefit abstract concepts. $x >= \sigma$ and $x <= -\sigma$ indicate highly concrete and highly abstract words accordingly

(a) abstract subset ($x < -\sigma$)

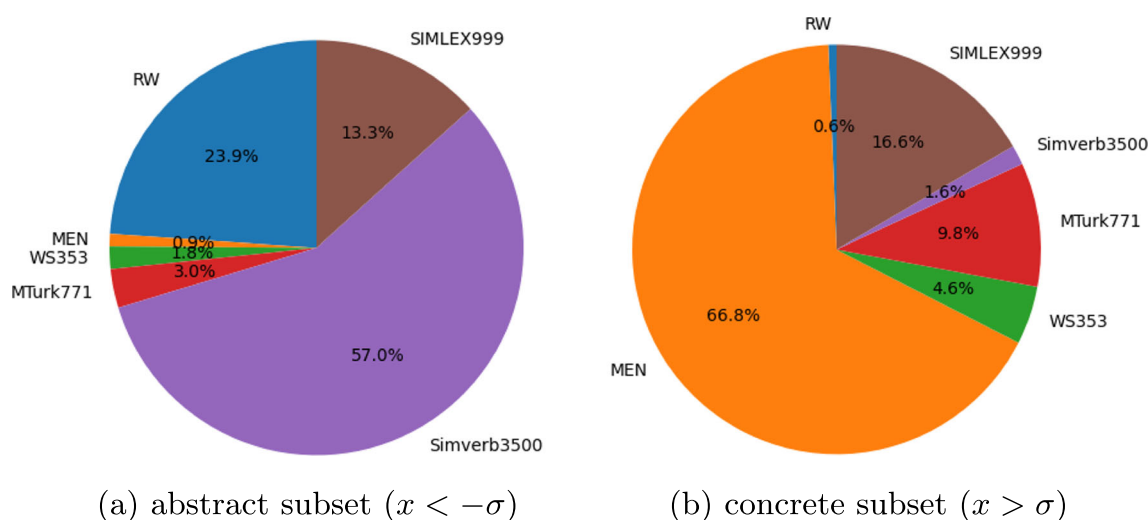(b) concrete subset ($x > \sigma$)

**Fig. 4** Dataset proportions for the highly abstract and highly concrete subsets of word pairs

and concrete words. To accomplish this, we devised a new experiment that draws upon the *WCR* dataset.

**Concreteness vs Abstractness** We computed a similarity score between each pair of the WCR dataset by applying the cosine similarity to the corresponding word vectors and used the Spearman correlation as the evaluation metric. We evaluated both the textual (GloVe) and visually grounded embeddings on four distinct subsets of the WCR with different concreteness scores. Concreteness subsets are obtained by the following steps.

1. To account for variations in concreteness scores, a standardization procedure is applied whereby scores are transformed into a standard normal distribution. Specifically, this involves subtracting the mean from all scores and dividing by their standard deviation, resulting in a standardized score $x_{is}$ for the $i$th word pair, expressed as $x_{is} : \frac{x_{in} - \mu}{\sigma}$.

2. After standardization, the distribution is partitioned into four segments based on the standard deviation and mean values, namely $[-\sigma, \mu, \sigma]$. The placement of word pairs within these segments allows for the differentiation of concrete and abstract word pairs. Specifically, pairs with higher concreteness scores are more likely to fall on the right side of the distribution ($x > \mu$), while those with lower scores are more likely to be located on the left side of the distribution ($x < -\mu$).

Results are shown in Fig. 3. Visual grounding leads to improved quality of textual embeddings regardless of the degree of concreteness. While the embeddings capture the meanings of concrete words more accurately in general, the

improvement is more significant for highly abstract words ($x <= -\sigma$). To investigate the potential cause of higher improvements for abstract words, we plotted the datasets' proportions of highly concrete words and highly abstract words in Fig. 4. Highly abstract word pairs are dominated by the *SimVerb3500* dataset, which seems to be the hardest for the textual embeddings to model (see Table 1). Highly concrete word pairs on the other hand mostly originate from the *MEN* benchmark, perhaps unsurprisingly, as it was compiled from image labels. The textual embeddings perform the best on this benchmark. Our finding is in line with previous works indicating that the meaning of concrete words is more stable and reliable compared to abstract words across different textual word embeddings (Pierrejean and Tanguy, 2019).

**Concreteness separation** Thus far, our findings demonstrate that the use of visual grounding leads to an improvement in the quality of embeddings for both concrete and abstract words. It is reasonable to assume that this is due to the grounding process creating a clearer separation between these two types of words. We carried out the following experiments to see whether this hypothesis holds true. We conducted training and assessment of two regression models by employing tenfold cross-validation on the MTurk40k dataset, which is a concreteness rating dataset assembled by Brysbaert et al (2014). The models utilized in this experiment included a straightforward linear regression and a multi-layer perceptron (MLP). The architecture of the MLP incorporated two hidden layers with 512 and 100 neurons, respectively. The models were given word representations as input and trained to predict the standardized concreteness scores.[8] Addition-

---

[8] https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html

ally, batch normalization (Ioffe and Szegedy, 2015) and dropout (Srivastava et al, 2014) techniques were integrated into the MLP model for better generalization. Dropout is a regularization technique to prevent overfitting by randomly dropping out (setting to zero) some neurons during training. Batch normalization improves the stability and speed of training by normalizing the inputs to each layer. Reported in Table 3, the difference between GloVe and our grounded embeddings (ZSG-G) is very subtle. This shows that visual grounding, as implemented in our model, does not necessarily cause stronger discrimination between concrete and abstract words.

**Nearest neighbors** For further exploration, we juxtaposed a sample of differing nearest neighbors of our best embeddings (ZSG-G) with its purely textual version (GloVe). Figure 1 shows the results for two random samples of highly abstract and highly concrete words in SimLex999. While GloVe retrieves related words (shown on the left), our grounding shifts the focus toward similarity and retrieves highly similar words for both concrete and abstract queries (shown on the right). We can observe that GloVe suffers from a bias toward the dissimilar words that frequently co-occur such as (many, people) and (sorta, weird). Our embeddings, on the other hand, alleviate this bias by creating more refined clusters of words. Even though our alignment is trained with mostly concrete words, the resulting vector space also benefits abstract words. In other words, abstract words are grounded indirectly via a learned mapping trained with concrete words. These findings align with the perspective of indirect grounding, which posits that concrete words are directly grounded while abstract words are indirectly grounded through language (Howell et al, 2005; Louwerse, 2011; Hoffman et al, 2018). Indirect grounding of abstract words has recently shown promising results in predicting abstract concepts using distributional semantic models (Utsumi, 2022).

Moreover, different typos of the same word such as 'peope' and 'poeple' (for people) occur with different frequencies in different contexts. Therefore, they are gradually pulled apart. Our model, however, puts them back into the same vicinity of space by applying the learnt alignment.

**Table 3** Mean Spearman's correlation coefficient $\times 100$ on MTurk40k using tenfold CV

| Model | GloVe tenfold score | ZSG-G tenfold score |
| --- | --- | --- |
| Linear regression | 84.90 | 84.70 |
| Multilayer perceptron | 88.86 | 88.24 |

Visually grounded embeddings (ZSG-G) do not seem to separate concrete and abstract words better in comparison to textual embeddings (GloVe)

## Alignment vs Fusion

In this and the subsequent section, we will conduct new experiments that manipulate the relationship between language and vision. These experiments will contribute to gaining deeper insight into the second question raised: how might language and embodied experiences work together to shape our comprehension of words? As the first step, various scenarios in which visual information could enhance textual word vectors are explored. In other words, we are interested to see whether increasing the influence of images on word vectors results in better grounded word vectors. For this aim, we train our model (ZSG-G) with different activation functions for the mapping $M$. Using a non-linear activation function such as ReLU and Leaky-ReLU (Xu et al, 2015) and adding more non-linear layers will allow the model to drastically deform the textual vector-space beyond linear transformations, increasing the influence of images on grounded word vectors. Table 4 shows the results with different numbers of layers and non-linear activation functions. We measure similarity and relatedness by evaluating on MTurk771 and SimLex999, as they are compiled for similarity and relatedness, respectively. Leveraging from different categories in SimLex999, we also evaluate on highly abstract and highly concrete words. Furthermore, for each case, we evaluate the obtained word vectors on all of the available datasets mentioned in Table 1. As shown in Table 4, we observe a consistent pattern of losing abstractness and gaining concreteness when non-linear transformations are used. This is to be expected since word vectors are morphing into image vectors and hence gain concrete properties. Employing two consecutive Leaky-ReLU is a prominent example of this case. Results on similarity and relatedness show that visual grounding shifts the focus toward similarity (see also Fig. 1). However, both similarity and relatedness are improved compared to textual embeddings by using a linear transformation, which helps benefiting from vision while keeping the textual information preserved. Overall, the best results on all the datasets are achieved by the linear mapping. This suggests that while visual information is beneficial for enhancing textual embeddings, giving too much emphasis to vision and neglecting language is not the optimal approach. These findings support previous evidence from case studies, as well as behavioral and neural studies, which suggest that abstract and concrete words are processed differently and involve distinct but overlapping brain regions (see Montefinese, 2019; Mkrtychian et al, 2019 for reviews). Therefore, it is crucial to strike a balance between concreteness and abstractness, which are represented in our experiments by visual properties of images and statistics of textual corpora, respectively. Language seems to benefit from vision the most when it is aligned/informed with vision as opposed to being completely fused together.

**Table 4** The impact of various activation functions and the number of layers used for the mapping $M$

| Type-Act.(No. of Layers) | Relatedness | Similarity | Abstract | Concrete | All |
|---|---|---|---|---|---|
| Textual GloVe | 71.5 | 43.3 | 43.3 | 40.2 | 56.7 |
| Grounded-Linear(1) | **73.2** | 52.6 | **54.6** | 59.2 | **63.6** |
| Grounded-ReLU(1) | 69.2 | 50.1 | 49.4 | 60.5 | 59.7 |
| Grounded-Leaky-ReLU(1) | 73.0 | **53.9** | 52.8 | 61.7 | 63.0 |
| Grounded-Leaky-ReLU(2) | 71.3 | 52.4 | 49.6 | **64.6** | 61.7 |

on-linear transformations led to a reduction in abstract knowledge and an increase in concreteness. The term "All" refers to the average score across all datasets listed in Table 1

As the first step, various scenarios in which visual information could enhance textual word vectors are explored. In other words, we are interested to see whether increasing the influence of images on word vectors results in better grounded word vectors. For this aim, we train our model (ZSG-G) with different activation functions for the mapping $M$. Using a non-linear activation function such as ReLU and Leaky-ReLU (Xu et al, 2015) and adding more non-linear layers will allow the model to drastically deform the textual vector-space beyond linear transformations, increasing the influence of images on grounded word vectors. Table 4 shows the results with different numbers of layers and non-linear activation functions. We measure similarity and relatedness by evaluating on MTurk771 and SimLex999, as they are compiled for similarity and relatedness, respectively. Leveraging from different categories in SimLex999, we also evaluate on highly abstract and highly concrete words. Furthermore, for each case, we evaluate the obtained word vectors on all of the available datasets mentioned in Table 1. As shown in Table 4, we observe a consistent pattern of losing abstractness and gaining concreteness when non-linear transformations are used. This is to be expected since word vectors are morphing into image vectors and hence gain concrete properties. Employing two consecutive Leaky-ReLU is a prominent example of this case. Results on similarity and relatedness show that visual grounding shifts the focus toward similarity (see also Fig. 1). However, both similarity and relatedness are improved compared to textual embeddings by using a linear transformation, which helps benefiting from vision while keeping the textual information preserved. Overall, the best results on all the datasets are achieved by the linear mapping. This suggests that while visual information is beneficial for enhancing textual embeddings, giving too much emphasis to vision and neglecting language is not the optimal approach. These findings support previous evidence from case studies, as well as behavioral and neural studies, which suggest that abstract and concrete words are processed differently and involve distinct but overlapping brain regions (see Montefinese, 2019; Mkrtychian et al, 2019 for reviews). Therefore, it is crucial to strike a balance between concreteness and abstractness, which are represented in our experiments by visual

properties of images and statistics of textual corpora, respectively. Language seems to benefit from vision the most when it is aligned/informed with vision as opposed to being completely fused together.

## Bridging the gap between language and vision

While our model is relatively simple compared to many others (Shahmohammadi et al, 2021; Kiros et al, 2018; Kiela et al, 2018), there are alternative approaches that use even simpler methods to integrate language with vision (Collell Talleda et al, 2017; Günther et al, 2022; Hasegawa et al, 2017). This raises the question of how to properly fill the gap between language and vision. We therefore investigated different ways in which the part of our model that bridges this gap can be engineered, and evaluated how well these alternative implementations perform.

We constructed the following scenarios. In all the scenarios, similar as before, after the training, we use the trained mapping $M$ to map all the textual embeddings into the grounded space to obtain grounded embeddings.

**Word-Level (WL)** For each training (caption, image vector) pair $(S_j, I_j) \in D$, we remove the stop words in caption $S_j$ and train a linear mapping $M$ from each word to its corresponding image vector $I_j$. For instance, the caption *'there is a dog on the floor'* would be converted into *'dog floor'*. Then, the textual embeddings of both *dog* and *floor* are mapped to their corresponding image one by one using only the mapping $M$. Similar to Günther et al (2022), we employed PCA (Pearson, 1901) to match the dimensions of the image vectors (2048) to the output of the mapping $M$ (1024).

**Bag-of-Words (BoW)** For each training (caption, image vector) pair $(S_j, I_j) \in D$, after mapping all the words in $S_j$ into the grounded space using a linear mapping here denoted again as $M$, we average them to obtain the BoW sentence representation. The BoW vector is then mapped into the image vector $I_j$ using a hidden layer with $Tanh$ activation function. This approach represents a more sophisticated method

than the 'Word-Level' model, as it utilizes all words in the captions and incorporates a non-linear transformation, potentially leading to improved performance.

**GRU** This set-up is very similar to our proposed model (see Section 2), and differs in that a single-layer GRU (Cho et al, 2014a) is used instead of an LSTM. A GRU is less complex compared to an LSTM and contains only a hidden-state as opposed to the LSTM, which is equipped with both a cell-state and a hidden-state.

**LSTM** This refers to the model proposed in Section 2.

**Transformer-Encoder (TE)** Attention-based sequence encoders introduced in Vaswani et al (2017) are currently used in state-of-the-art contextualized language models (Lan et al, 2019; Devlin et al, 2018) and are applied to complex downstream NLP tasks. We are interested in whether the utilization of cutting-edge NLP techniques can enhance the capacity to capture human-rated word similarity and relatedness. These encoders generate contextualized embeddings based on the learnable associations between words, allowing for the disambiguation of polysemous words in different contexts. For instance, the word *'clip'* has different senses in *'I clip my nails'* and *'I saw a video clip'*. To distinguish between these senses, contextualized representations of *'clip'* are therefore computed that are informed by its associations with the words in a given context. For our experiments, we pass the textual embeddings of each caption through the mapping $M$ as before. Then we train a different number of encoders on top of $M$. That is, the embeddings are passed through multiple transformer encoders simultaneously. The output of the encoders is the contextualized representation of the given caption which is then projected to the image vector through a linear layer. We constructed the transformer encoders with 1024 hidden size, 16 atten-

tion heads and used NAdam with the learning of 0.0001 for training. For a comprehensive understanding of transformer architecture, we highly recommend referring to the seminal work by Vaswani et al (2017).

The results of each model configuration are reported in Table 5. Notably, the word-level mapping fails to preserve a sufficient amount of textual information, resulting in embeddings that are significantly distorted when compared to text-only embeddings. As a consequence, these embeddings demonstrate inferior performance across all datasets. We note here that a single image is very rich in information and often is not well described by a single word. Furthermore, the relationship between language and vision is not always linear or straightforward. For instance, many highly concrete nouns and adjectives such as *apple* and *red* could be easily coupled with their visual representations. In contrast, more abstract linguistic categories such as prepositions and conceptual words establish their link to visual experiences through intricate (not necessarily linear) statistical patterns embedded within language.

While the BoW model does offer some improvement over the text-only GloVe approach on certain datasets, its overall performance is relatively comparable. However, it is worth noting that the BoW model demonstrates significant enhancement on the SimLex999 dataset, which evaluates word similarity rather than relatedness. Conversely, its performance is weaker on the MTurk771 dataset, which focuses on relatedness. The potential reason for these fluctuations in performance is that the BoW representations do not account for word order and, consequently, lose the temporal statistics of how related words co-occur within their context (see Jones and Mewhort, 2007, for embeddings jointly representing word meaning and word order).

The utilization of recurrent neural networks (specifically, GRU and LSTM models) results in significantly improved

**Table 5** Evaluation of various textual encoders reveals a consistent improvement in performance from the most simplistic approach (WL) to the utilization of an LSTM model

| 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 |
| WL | 27.7 | 49.7 | 34.2 | 31.7 | 7.10 | 1.50 | 25.3 |
| BoW | 46.5 | 75.2 | 73.8 | 60.1 | 33.8 | 46.0 | 55.9 |
| GRU | 51.2 | 83.0 | 75.1 | 71.3 | 36.9 | 48.3 | 60.1 |
| LSTM | **53.4** | **85.1** | **78.8** | **73.2** | **38.5** | **52.6** | **63.6** |
| 1-layer-TE | 44.0 | 77.4 | 62.9 | 67.0 | 25.5 | 37.5 | 52.9 |
| 2-layer-TE | 50.1 | 82.6 | 75.3 | 72.2 | 32.4 | 45.6 | 59.7 |
| 3-layer-TE | 50.0 | 82.0 | 72.7 | 72.2 | 33.0 | 46.8 | 59.4 |

However, In light of our experimental results, it appears that transformer-encoders may not be particularly well suited for generating visually grounded word embeddings

performance. Of these two models, the LSTM outperforms the GRU, which is unsurprising given its ability to effectively capture long-distance dependencies between words and encode the entirety of a sentence.

However, training with a single transformer encoder fails to produce better quality embeddings, perhaps unsurprisingly as these encoders are usually stacked on top of each other to achieve the desired outcome (Vaswani et al, 2017). We therefore also tested models with two and three layers of TE. While using a two-layer TE demonstrated improved performance, we did not observe any further improvement with additional layers beyond that. We also employed multiple layers of LSTM and found that a single-layer LSTM produces the most favorable outcomes. While adding more layers typically results in a more robust model, we contend that as the network grows deeper, there is a decreased amount of visual knowledge that can be easily conveyed back to the mapping $M$. In other words, the visual knowledge becomes distributed across various layers, making it arduous to distill the information down into a single layer. Recall that after the training we only use the mapping $M$ to obtain visually grounded representations. Consequently, a network that effectively condenses information within $M$ while accurately predicting image vectors is highly desirable. In our experiments, we found that a single-layer LSTM strikes the ideal balance between the degree of dependence on $M$ and producing high-quality image vectors.

In summary, our experiments in the last two sections aimed to apply computational models to shed light on the question of how language and embodied experiences (here crudely represented as images) might interact to shape our comprehension of words. In our experiments, a linear transformation in isolation is not adequate for establishing a strong connection between vision and language. In order to obtain high-quality visually grounded embeddings, it is imperative to incorporate a non-linear transformation. Furthermore, it is essential to carefully calibrate the semantic space of the textual embeddings to accurately capture the perceptual knowledge present in images. Allowing too much influence from the visual modality may lead to distortion of the textual embeddings, emphasizing the importance of striking a delicate balance between the two modalities. This finding suggests that also the human mind integrates information from vision in its semantic system, but that this system is not dominated by visual similarities. It is worth noting that philosophers such as Kant, Husserl, and Merseau-Ponty have pointed out that we do not perceive the world as it truly is, our perceptions are shaped by our senses, the constraints imposed by the world on our survival, and our cultures (see, e.g., Kant et al, 1999; Husserl, 1913; Merleau-Ponty et al, 2013). A very similar point was made more recently from the perspective of the cognition of vision by Hoffman (2019). The way in which we implement visual grounding—constraining the extent to which vision can change embeddings from human texts—does justice, however crude, to this fundamental insight.

## Contextualized visual grounding

While we successfully showed the benefit of visual grounding for word embeddings on a wide range of intrinsic tasks, it remains a topic of debate as to whether visual grounding provides benefits for state-of-the-art NLP models on sentence-level language tasks (Yun et al, 2021; Iki and Aizawa, 2021; Tan and Bansal, 2020). While some recent approaches have reported minor improvements through the use of visually grounded models (Sileo, 2021), there is a growing consensus that these models, such as VL-BERT (Su et al, 2019), do not provide significant benefits for language tasks. In fact, there is concern that these models may distort the linguistic knowledge acquired from textual corpora and hinder their effectiveness for natural language understanding tasks (Tan and Bansal, 2020; Yun et al, 2021) and modeling abstract concepts (Pezzelle et al, 2021). Currently, transformers have achieved state-of-the-art performance on a wide range of downstream NLP tasks. Transformers are a type of deep contextualized language model that typically operate using stacked attention layers. These models are capable of capturing long-range dependencies in language by attending to relevant words in the input sequence at each layer, allowing them to achieve impressive performance on a variety of NLP tasks (Vaswani et al, 2017) (briefly explained in Section 6). Many of these models, such as BERT (Devlin et al, 2018), undergo a two-phase process, consisting of pretraining and fine-tuning. During pretraining, the model is trained on a masked language modeling task, whereby certain tokens within the input sequence are masked, and the model is trained to predict the masked tokens. This process enables the model to acquire a deep understanding of the underlying linguistic structure of the language, including its syntax and semantics. In the subsequent fine-tuning phase, the pretrained model is further optimized for performance on downstream tasks, such as sentiment classification (Socher et al, 2013) and paraphrase detection (Dolan and Brockett, 2005). By fine-tuning the model on these specific tasks, it can be tailored to achieve state-of-the-art results, leveraging the powerful contextualization capabilities of the transformer architecture. For instance, in the case of sentiment classification, a new multi-layer perceptron (MLP) could be appended to the encoded output of the main model to generate a binary decision for a given sentence. The parameters of both the added MLP and the pretrained model can then be fine-tuned using the available training data for sentiment classification. With the abundance of training data, the vast amount of textual context, and the powerful capabilities of the transformer architecture, one could argue that visual grounding does not
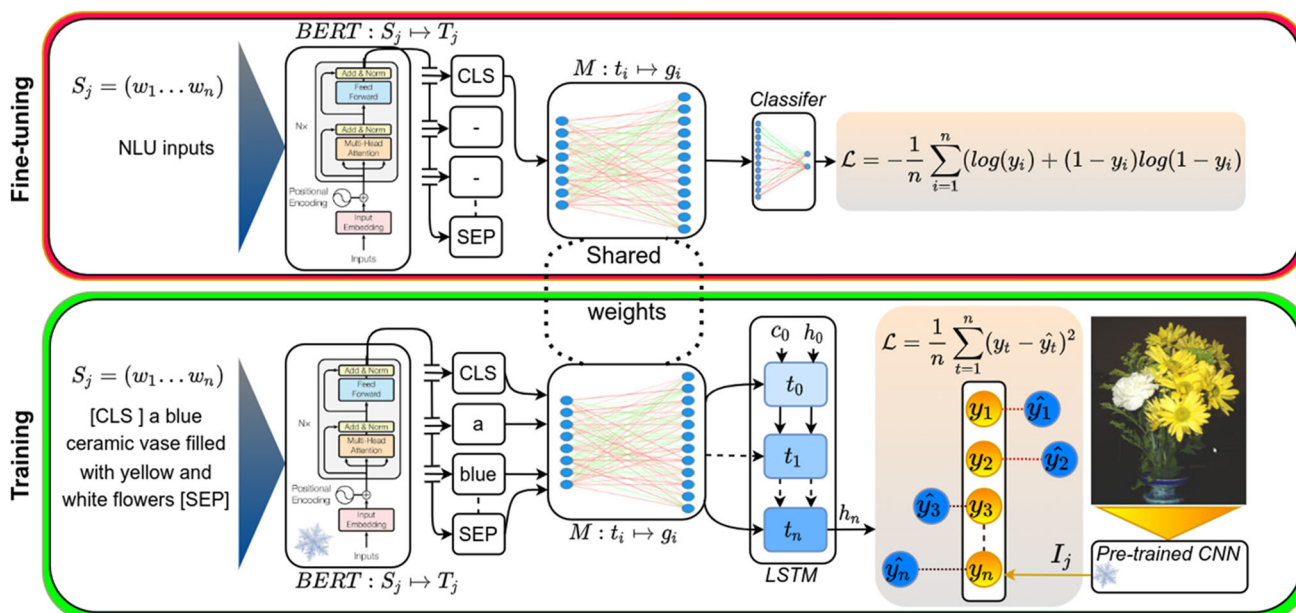
**Fig. 5** We construct a visually grounded version of BERT using image–caption pairs. In the training phase, the frozen pre-trained BERT encodes the caption, and an alignment $M$ followed by an LSTM layer on top of BERT is trained to predict the corresponding image vector. In the fine-tuning phase, the learned alignment $M$ is attached on top of BERT followed by a classifier. This alignment ensures that the BERT representations are guided by the learned visual alignment during fine-tuning

offer any additional information for solving current NLP tasks (Tan and Bansal, 2020).

Despite the arguments against the necessity of visual grounding for transformer-based language models, we are curious about the potential benefits of our simple grounding approach. To explore this possibility, we incorporated our approach with BERT (Devlin et al, 2018), one of the pioneering transformer models for sentence-level natural language understanding tasks. BERT has been pre-trained on a vast corpus of English text, including English Wikipedia[9] and BookCorpus (Zhu et al, 2015), a collection of 11,038 unpublished books. We carry out new experiments to compare the performance of visually grounded BERT and purely textual BERT on sentence-level NLP tasks. To clarify, in our baseline model, fixed FastText or GloVe vectors serve as the input to the **M** mapping. However, in our new model, these vectors are replaced by vectors generated through BERT encoding. The BERT encoder marks the beginning and end of the input with '[cls]' and '[sep]' tokens (as shown in Fig. 5) and outputs a fixed-dimensional vector for each token. Therefore, we can treat it as a word-embedding model. Given a sentence ($S_j = [w_1, w_2, \cdots, w_n]$) with $n$ words, the BERT encoder outputs ($T_j = [t_1, t_2, \cdots, t_n]$), where $t_i$ represents the contextualized encoding of the word $w_i$.

When used for classification tasks, the BERT engine is coupled with a multi-layer-perceptron network generating the final output. As shown in Fig. 5, similar to our proposed model, we train a linear mapping $M$ followed by an LSTM encoder to predict an image vector given its caption. After the training phase (see the lower box), for each classification task, the pre-trained model has to be fine-tuned. For this step, an MLP is added on top of the mapping $M$ for fine-tuning on the downstream task (see the upper box). In the fine-tuning phase, the '[cls]' tokens encode the given input through multiple attention layers and the rest of the tokens are discarded (Devlin et al, 2018). In a nutshell, our approach adds the learned alignment $M$ between the pre-trained BERT encoder and its classifier. This alignment is applied to the BERT encoding to align its final representation to vision without deteriorating its textual information.

**Evaluation** We fine-tuned and evaluated our pre-trained grounded BERT on the General Language Understanding Evaluation (GLUE) benchmark[10] (Wang et al, 2018) implemented in the Huggingface[11] library (Wolf et al, 2019). GLUE is widely regarded as a comprehensive evaluation suite for natural language understanding models that reflect a wide range of the complexity and diversity of human language comprehension. It consists of nine natural language understanding tasks: single-sentence tasks, SST-2 (Socher et al, 2013) and CoLA (Warstadt et al, 2019); paraphrasing and similarity tasks, MRPC (Dolan and Brockett, 2005),

---

[9] https://en.wikipedia.org/wiki/English_Wikipedia

[10] https://gluebenchmark.com/

[11] https://huggingface.co/

QQP[12], and STS-B (Cer et al, [2017]); natural language inference tasks, RTE (Wang et al, [2018]), QNLI (Rajpurkar et al, [2016]), MNLI (Williams et al, [2017]), and WNLI (Levesque et al, [2012]). In what follows, we briefly explain the GLUE tasks used in our experiments.

**SST-2** The Stanford Sentiment Treebank compiles a set of sentiment annotations from movie reviews. It includes a total of 215,154 phrases each annotated by three human annotators. Each sample is assigned to one of the following five labels: neutral, slightly neutral, moderately positive, or positive. SST-5 or SST fine-grained refers to the corpus with all five labels. SST-2, however, consists of binary labels only. The negative class indicates negative or slightly negative and the positive class indicates somewhat positive or positive. The neutral sentences are discarded in SST-2 resulting in 70,042 overall samples. Examples of positive and negative sentences are *'that loves its characters and communicates something rather beautiful about human nature'* and *'that 's far too tragic to merit such superficial treatment'* accordingly.

**CoLA** The Corpus of Linguistic Acceptability is an English acceptability evaluation dataset. It consists of 10,657 sentences from 23 linguistics publications, expertly annotated for acceptability (grammaticality) by their original authors into positive and negative classes. Some negative examples are: *'The professor talked us'*, *'They made him to exhaustion'*, and *'The witch went into the forest by vanishing'*.

**MRPC** The Microsoft Research Paraphrase Corpus is a set of sentence pairs retrieved from online news sources. MRPC includes 5801 sentence pairs, each labeled by human judges as to whether the pair constitutes a paraphrase. This task is also known as paraphrase detection. Examples from this dataset are, **positive:** (*'About 130,000 U.S. troops remain in Iraq , with others deployed in Afghanistan, South Korea and elsewhere.'*, *'About 130,000 US soldiers remain in Iraq , with others serving in Afghanistan, South Korea, Japan, Germany, and elsewhere.'*); **negative:** (*'The Embraer jets are scheduled to be delivered by September 2006.'*, *'The Bombardier and Embraer aircraft will be delivered to U.S. Airways by September 2006.'*).

**QQP** The Quora Question Pairs, is a collection of question pairs from the question-answering website Quora. The task is identical to that of MRPC. the QQP, however, is much larger, it compiles a set of 400,000 question pairs each with a binary label indicating the semantic equivalence of the question pair.

---

[12] https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs

**STS-B** The Semantic Textual Similarity Benchmark is a set of sentence pairs compiled from captions for videos and images, natural language inference data, and news headlines. It consists of 8628 sentence pairs with each pair annotated by humans with a similarity score ranging from 1 to 5. The task is to predict the similarity score of a given pair as a real-valued number. For example, (*'A woman is dancing.'*, *'A man is talking'*) has a score of 0 and (*'A small dog is chasing a yoga ball'*, *'A dog is chasing a ball'*) has a score of 4.

**RTE** Recognizing Textual Entailment is the task of modeling a directional relation between two sentences. The relation holds whenever the truth of the second sentence is entailed by the first one. For instance, *'a dog is jumping for a Frisbee in the snow'* entails *'An animal is outside in the cold weather, playing with a plastic toy.'* but contradicts *'a cat washed his face and whiskers with his front paw.'*. The RTE dataset consists of 5767 pairs, extracted from news and Wikipedia text, each with a binary label.

**QNLI** The Stanford Question Answering Dataset consists of question-paragraph pairs. One of the sentences in the paragraph (drawn from Wikipedia) contains the answer to the question in the given sample. Questions are written by human annotators. To convert this task into a sentence pair classification one, Wang et al ([2018]) constructed a pair between each question and each sentence in the corresponding paragraph, and discarded pairs with low lexical overlap between the question and the context (paragraph) sentence. The task is to predict whether the context sentence contains the answer to the question. This dataset contains 115,699 question-sentence pairs each annotated with a binary label. Examples from this dataset are, **positive:** (*'When is the term 'German dialects' used in regard to the German language?'*, *'When talking about the German language, the term German dialects is only used for the traditional regional varieties.'*), **negative:** (*'In what century was the church established at the location?'*, *'Construction of the present church began in 1245, on the orders of King Henry III.'*)

**MNLI** The Multi-Genre Natural Language Inference is a dataset of 431,992 sentence pairs with entailment annotations. Given a pair of premise-hypothesis sentences, the task is to predict whether the premise entails the hypothesis (entailment), contradicts the hypothesis (contradiction), or neither (neutral). The premise sentences are gathered from different sources including government reports, transcribed speech, and fiction. There are two versions of the validation set, matched and mismatched. The former contains samples in the same domain as in the training set while the latter contains cross-domain samples. We evaluate our model on both sets.

**Table 6** Validation scores on the GLUE benchmark using textual BERT and visually grounded BERT (*_GBERT)

| Model/Data | CoLA | MRPC | QNLI | QQP | RTE | SST-2 | MNLI | STS-B | Mean Score |
|---|---|---|---|---|---|---|---|---|---|
| Train Size (K) | 8.5 | 3.6 | 104 | 364 | 2.5 | 67 | 392 | 5.7 | - |
| Textual-BERT | 59.05 | 84.31/89.15 | 91.08 | 90.76/87.53 | 67.15 | 91.2 | 83.34/83.83 | 87.13/87.00 | 81.74 |
| 1-LFM-GBERT | 60.07 | 84.31/89.11 | 91.00 | 90.82/87.63 | 63.54 | **92.43** | 83.86/83.52 | 88.83/88.49 | 81.86 |
| 2-LFM-GBERT | 61.58 | 85.29/89.58 | 91.47 | 90.71/87.44 | 67.15 | 92.09 | 83.78/83.66 | 88.44/88.04 | 82.56 |
| 3-LFM-GBERT | 60.62 | 84.56/89.44 | 90.92 | 90.70/87.46 | **68.23** | 92.32 | 83.84/83.48 | 88.02/87.67 | 82.40 |
| 2-LTM-GBERT | **61.62** | **86.27/90.51** | 91.12 | 90.73/87.46 | 67.15 | 92.20 | 83.73/83.71 | **89.12/88.74** | **82.74** |

Visual grounding seems to improve the generalization of the model when training data is limited (e.g., MRPC and CoLA). However, large volumes of training data compensate for visual grounding (see the scores of QQP and MNLI). *accuracy/F1_scores* are reported for QQP and MRPC, *Pearson/Spearman* correlations are reported for STS-B, and accuracies for *matched/mismatched* sets are reported for MNLI. For the other tasks, accuracy is reported. Numbers in bold indicate obvious improvements over textual BERT

**Implementation details** We used the *bert-base-cased* version of BERT (Devlin et al, 2018) in our experiments. '*base*' refers to the size of the model in terms of the number of training parameters. There are three versions of BERT: *small, base,* and *large*; 'cased' indicates that the model distinguishes between upper-cased and lower-cased letters. For training, we used the Microsoft COCO 2017 dataset (Lin et al, 2014). The alignment $M$ maps a BERT token $t_i \in \mathbb{R}^{768}$ to $g_i \in \mathbb{R}^{1024}$. Each LSTM layer contains 1024 units. A single-layer neural network with a linear activation function (a linear layer) is applied on top of the LSTM to predict the image vector $I_j \in \mathbb{R}^{2048}$. We trained the model on image–caption pairs for ten epochs using the AdamW optimizer (Loshchilov and Hutter, 2017) with the learning rate set to $5e^{-5}$ and a batch size of 64. For fine-tuning on the GLUE benchmark, we followed the huggingface guidelines[13] and fine-tuned the model on each downstream task for five epochs with a batch size of 32 and a learning rate of $2e^{-5}$.

**Results** Table 6 reports the validation scores across the GLUE datasets. The WNLI dataset was excluded from the list following Devlin et al (2018) due to inconsistent results. We carried out our grounding experiments with different numbers of LSTM layers. In Table 6, *n-LFM-GBERT* indicates the grounded BERT with *n* layers of LSTMs and frozen (weights are kept unchanged during training) mapping **M** while fine-tuning on downstream tasks. The idea behind freezing the mapping (alignment) **M** while fine-tuning the BERT encoder and the classifier on a particular task is to guide (force) the output representations of BERT to follow the visual alignment. This might then guide the model to a better feature space for solving the task. Considering the mean score, the grounded model with 2-layer-LSTMs (*2-LFM-GBERT*) outperforms the textual BERT by almost 1%, highlighting the potential benefits of visual grounding. Moreover, we also fine-tuned the alignment **M** of the best model

---

(*2-LFM-GBERT*) for each particular task along with BERT encoder and the classifier, denoted as *2-LTM-GBERT*, this model further improves the results. Although the improvements achieved through visual grounding in our experiments are marginal compared to those obtained through grounded word embeddings, the results presented in the table provide valuable insights. Notably, for datasets with limited training data, such as CoLA and MRPC, visual grounding appears to provide an advantage, as indicated by the bold numbers in the table. However, for larger datasets such as QQP and MNLI, the results are almost identical for both grounded and textual BERT models. These findings suggest that visual grounding improves the generalization of transformers when training data is limited. Nonetheless, they also demonstrate that a substantial amount of textual training data, combined with meticulous fine-tuning of models, can compensate for the relatively simple visual grounding approaches used in our study when tested on the GLUE benchmark. In accordance with our prior word embeddings experiments, we conducted a *t* test comparing the results of *textual BERT* to those of *Grounded BERT*, more specifically *2-LTM-GBERT*. The statistical test indicated that the observed enhancements in performance were **not** statistically significant. Nevertheless, when compared to the process of human language acquisition, these textual language models exhibit significant inefficiencies, requiring exposure to vast amounts of training data and computational resources to achieve satisfactory results (Strubell et al, 2019). The BERT model for instance, despite being pre-trained on an extensive corpus of over 3 billion tokens, still requires meticulous fine-tuning for each individual task, which raises doubts about the efficacy of large language models and the potential usefulness of visual grounding in this regard.

In light of these concerns, we conducted an investigation to determine whether fine-tuning the model would obscure the benefits/improvements in the overall quality of embeddings due to visual grounding. In other words, fine-tuning the models might diminish the differences between them, as the

learned parameters are tailored to the specific downstream task, potentially obscuring the benefits of visual grounding. For this aim, we designed a new experiment whereby we skipped the fine-tuning phase and conducted a comparative analysis of the semantic spaces of Textual BERT and Grounded BERT models. Despite the adverse impact of skipping fine-tuning on the results, this experimental approach enables us to juxtapose the semantic space of the two models more accurately and identify potential subtle differences between them, with a particular focus on the influence of visual grounding. To compare the semantic space of Grounded BERT and Textual BERT for each specific task within the GLUE benchmark, we employ a technique called *linear probing*. In this technique, only a linear classifier such as logistic regression is trained on top of pre-trained representations of a model, in order to measure the quality of the learned representations for particular downstream tasks (Reif et al, 2019). For tasks involving pairs of sentences, a linear probe is trained with the cosine similarity between the representations of the two sentences. For instance, consider the task of paraphrase detection using the MRPC dataset, which involves predicting whether a given pair of sentences is semantically equivalent. In our probing setup, the two sentences, $s_1$ and $s_2$, are first encoded separately by Grounded BERT and Textual BERT, resulting in two vectors, $v_1$ and $v_2$, representing each sentence. We then determine the semantic similarity of the two sentences by calculating the cosine similarity between the two vectors:

$$score(v_1, v_2) = 1 - \frac{v_1.v2}{\|v_1\|\|v_2\|}.$$

After encoding the sentences and calculating cosine similarities between the two vectors, a logistic regression model (the probe) is trained using the cosine similarities as inputs and binary classification labels as outputs. Following training, the trained linear probe is applied to predict the labels of the validation set. The rest of the evaluation procedure is identical to the previous section. If one of the models' representations is better suited for this task, we expect to observe higher performance, indicating better classification boundaries and more refined clusters in the semantic space of the model.

The evaluation results of probing are reported in Table 7. Grounded BERT demonstrates significant improvements over textual BERT leading to the enhancement of the mean score by 5%. This shows that visual grounding enriches language representations across a wide range of abstract language understanding tasks. Surprisingly, the accuracy on *CoLA* dataset, is higher than when the whole model is fine-tuned (see Tabel 6). This might be due to the nature of the task. Since negative samples contain ungrammatical sentences, they might inherently be well separated from correctly grammatical sentences in the vector space. Hence, fine-tuning the parameters of BERT with a small set of ungrammatical sentences might be detrimental to model performance. This further confirms the inefficiencies of large language models and their need to devour a huge amount of annotated data to achieve desirable performance. We further performed a t-test between the prediction of the two models, exhibiting statistically significant differences between the performance of the two models on the majority of the tasks. Bold numbers in Table 7 indicate $p$ values $< 0.05$.

Overall, these insights highlight the potential of visual grounding even for highly advanced NLP techniques. Our findings suggest that visual grounding has the potential to learn task-agnostic language representations, leading to reduced computational costs and textual resources. This paves the way for future research on building cognitively plausible language learning frameworks where the learning process leverages different modalities such as visual cues and gestures (Smith and Gasser, 2005; Iverson and Goldin-Meadow, 2005), making the learning both effective and cognitively plausible.

## Grounding for smaller datasets

Thus far, our grounding approach has been shown to be effective in conjunction with pre-trained word embedding models and advanced sentence-level language models, when training data for a given downstream task is scarce. In both cases, however, large amounts of textual training data from different domains have been utilized. The amount of training data plays a big role in shaping performance on downstream tasks (Beltagy et al, 2019; Lee et al, 2020), and in general is an important determinant of the quality of industrial word embeddings (Wang et al, 2019; Elekes et al, 2018; Johns and Jones, 2022). This section details two concluding exper-

**Table 7** Validation scores on the GLUE benchmark by employing a linear probe on textual BERT and visually grounded BERT

| Model/Data | CoLA | MRPC | QNLI | QQP | RTE | SST-2 | MNLI | STS-B | Mean Score |
|---|---|---|---|---|---|---|---|---|---|
| Textual-BERT | 73.92 | 68.38/81.22 | 52.48 | 63.18/00.00 | **52.35** | 82.34 | 36.40 | 22.70/09.78 | 56.47 |
| Grounded-BERT | **77.85** | 68.38/81.22 | **54.42** | **67.21/48.63** | 48.38 | 85.55 | **42.25** | **47.80/47.29** | 61.48 |

The visually grounded vector space provides richer semantic representations, leading to improved language understanding on a majority of the tasks. Numbers in bold indicate significant differences in performance ($p$ values $< 0.05$)

**Table 8** Comparison of our grounded embeddings (*-G) to textual embeddings (*-T) on limited training data

| Model | RW | MEN | WSim 353 | MTurk 771 | SimVerb 3500 | SimLex 999 | Mean |
|---|---|---|---|---|---|---|---|
| TASA-T | 2.4 | 37.2 | 33.1 | 35 | 8.5 | 10.8 | 21.7 |
| TASA-G | **6.7** | **42.5** | **37.1** | **37.5** | **13.1** | **17.1** | **25.7** |
| Text8-T | 8.1 | 47.9 | 45.9 | 45 | 8.3 | 16.6 | 28.63 |
| Text8-G | **13.5** | **51.2** | **51.8** | **49.1** | **10.5** | **20.7** | **32.8** |

GloVe algorithm was trained on TASA and Text8 corpus separately from scratch. Significant improvements are achieved by visual grounding despite limited training data. Numbers in bold indicate significant differences in performance ($p$ values $\leq 0.0008$, $t$ test)

iments that address the question of whether visual grounding is also beneficial for embeddings calculated from much more modest training data. As human lexical acquisition develops rapidly on the basis of restricted amounts of training data, a solid improvement due to visual grounding even under limited exposure would provide support for the possibility that human learning also benefits from visual grounding.

We, therefore, trained the GloVe model from scratch on two small and different training corpora and measured the improvements of our grounding approach on each corpus using the word similarity benchmarks (see Section 4). Initially, we obtained textual embeddings by training on two distinct corpora: TASA and Text8. TASA (Zeno et al, 1995) has served as a training corpus for, e.g., Latent Semantic Analysis (Landauer, 1999). Text8 is a small corpus sampled from Wikipedia to allow quick testing of language models.[14] Our best grounding model (see Section 2) is then applied to the textual embeddings to obtain visually grounded embeddings. Table 8 reports the comparison between textual embeddings and grounded embeddings for both corpora. Our grounded approach (TASA-G and Text8-G) consistently improves on top of textual embeddings (TASA-T and Text8-T) despite the small size of these corpora and the very different nature of the training corpora. We further confirmed the statistical significance ($p \leq 0.0008$) of the performance improvements observed by conducting $t$ tests on both datasets. The robustness of our grounding method for word-based embeddings holds not only across a wide range of tasks, but also for different amounts of training data, providing a firm basis for expecting grounded embeddings to provide improved precision to studies of human cognition that make use of embeddings.

## Discussion and conclusion

In this study, we designed a visual grounding framework that effectively produces visually grounded word representations for all types of words from different kinds of embeddings.

Our approach, apart from its simplicity, shows excellent generalization, as evidenced by its success on a variety of human-annotated similarity and relatedness tasks, including those involving unseen abstract and concrete words. We have made both the grounded embeddings and our framework publicly available. We further designed a series of experiments to shed light on the following research questions.

**Visual grounding for abstract words** Our approach employs a visual grounding pathway that is acquired during the process of grounding concrete words, which enables the indirect grounding of abstract concepts. Our study's results lend support to the indirect grounding theory, which posits that concrete words are directly grounded while abstract words are indirectly grounded through language (Howell et al, 2005; Louwerse, 2011; Hoffman et al, 2018). Despite being trained on image captions within which concrete nouns far outnumber abstract nouns, our approach produces more refined clusters of both concrete and abstract words, highlighting the framework's ability to capture the subtle nuances in the semantics of different word types across a wide range of human-annotated word collections.

**Bridging language to vision** We investigated various strategies of bridging language (here crudely represented as word/sentence embeddings) with vision. Our experiments support the following conclusions.

First, textual word embeddings benefit from vision the most when they are aligned with vision as opposed to being merged. Our alignment strategy enables the textual embeddings to incorporate real-world knowledge through images without compromising the statistical knowledge gained from textual corpora. We showed by example that allowing too much visual information will overwhelm the textual embeddings. Injecting too much visual knowledge into the embeddings benefits concrete words while diminishing the performance on modeling abstract words. This trade-off may be due to the distinct cognitive processing of abstract and concrete words, which engage overlapping but separate brain regions (see Montefinese, 2019; Mkrtychian et al, 2019, for reviews). Therefore, the right balance between concreteness and abstractedness represented in our experi-

---

[14] https://cs.fit.edu/~mmahoney/compression/textdata.html

ments by visual properties of images and statistics of textual corpora is vital.

Our second key finding is that textual context plays an important role in grounding isolated word embeddings. Our results demonstrate that linking word embeddings with vision in the absence of textual context leads to a significant distortion of the semantic space. We believe one reason is that word vectors still need to be aware of the textual context they occur in when they are being coupled with their corresponding visual information in images. Moreover, given that images are a highly complex and rich source of information, a single word cannot capture their full semantic richness. Our grounding framework, therefore, aligns word vectors with their corresponding images while simultaneously preserving information about their textual context, thereby enhancing the overall efficacy of the grounding process.

**Benefits and upper bound of visual grounding** Our study has demonstrated that visual grounding is highly advantageous for both concrete and abstract words. However, our analyses have also revealed that visual grounding is particularly beneficial in cases where textual embeddings struggle, such as when modeling highly abstract verbs or rare words. Conversely, in benchmarks consisting mostly of concrete words, the improvement from grounding is less pronounced. These findings dovetail well with the observation that the meanings of concrete words are more stable and reliable compared to those of abstract words across different textual word embeddings (Pierrejean and Tanguy, 2019).

It has been demonstrated that infants' capacity to process abstract words develops later, after they have established a firm foundation in concrete concepts (Bergelson and Swingley, 2013, 2012). Furthermore, many abstract concepts build on metaphors that themselves are rooted in concrete experiences (Lakoff and Johnson, 1980; Langacker, 1987). This finding suggests a possible high-level explanation of why abstract words benefit from visual grounding of concrete words: Abstract words are scaffolded on the foundations of concrete words. Visual grounding contributes to a more precise approximation of these foundations, and this in turn enables a recalibration of the superstructure of abstract words. Our findings thus pave the way for future research on whether visual grounding alleviates the instability problem of abstract concepts (Pierrejean and Tanguy, 2019).

**Visual grounding and corpus size** The embeddings used in current NLP are derived from corpora comprising billions of words. An examination of the extent to which visual grounding helps improve state-of-the-art sentence-level NLP models built on such huge resources revealed only modest improvements. Specifically, a comparison of a visually grounded version of the well-known BERT model (Devlin et al, 2018) with a standard textual version of BERT on common evaluation benchmarks showed that visual grounding yields considerable improvements only when training data is limited. However, when using large volumes of textual data and meticulous parameter-tuning, the performance of the visually grounded and textual models becomes almost identical. Apparently, huge volumes of textual context in combination with subsequent powerful fine-tuning algorithms compensate for visual grounding, at least on current downstream NLP tasks.

Although visual grounding is not necessary for language models that have access to volumes of data that far surpass what individual speakers can ever encounter, we have shown that when embeddings are trained on small corpora, visual grounding leads to substantial improvements.

Since we as humans are never exposed to the amount of textual data digested by current language models, but still master our first language at a very early age, enriching current models for lexical semantics with vision is a first step forward in the direction of developing cognitively more plausible representations for word meaning.

## Open practices statement

Code for acquiring grounded word embeddings and two sets of ready-to-use grounded embeddings are available at https://github.com/Hazel1994/Visually_Grounded_Word_Embeddings_2.

# References

Abdou, M., Kulmizev, A., Hershcovich, D., et al. (2021). Can Language Models Encode Perceptual Structure Without Grounding? A Case Study in Color. In: *Proceedings of the 25th conference on computational natural language learning. association for computational linguistics*, Stroudsburg, PA, USA, pp. 109–132, https://doi.org/10.18653/v1/2021.conll-1.9

Anderson, A. J., Bruni, E., Lopopolo, A., et al. (2015). Reading visually embodied meaning from the brain: Visually grounded computational models decode visual-object mental imagery induced by written text. *NeuroImage, 120*, 309–322.

Andrews, M., Frank, S., & Vigliocco, G. (2014). Reconciling embodied and distributional accounts of meaning in language. *Topics in Cognitive Science, 6*(3), 359–370.

Avery, J. E., Goldstone, R. L., & Jones, M. N. (2021). Reconstructing maps from text. *Cognitive Systems Research, 70*, 101–108.

Baayen, R.H., Chuang, Y.Y., Shafaei-Bajestan, E., et al. (2019). The discriminative lexicon: A unified computational model for the lexicon and lexical processing in comprehension and production grounded not in (de) composition but in linear discriminative learning. *Complexity* 2019

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

Barsalou, L. W. (2008). Grounded Cognition. *Annual Review of Psychology,59*(1). https://doi.org/10.1146/annurev.psych.59.103006.093639

Barsalou, L. W. (2003). Abstraction in perceptual symbol systems. *Philosophical Transactions of the Royal Society B: Biological Sciences, 358*(1435), 1177–1187. https://doi.org/10.1098/rstb.2003.1319

Barsalou, L. W., Dutriaux, L., & Scheepers, C. (2018). Moving beyond the distinction between concrete and abstract concepts. *Philosophical Transactions of the Royal Society B: Biological Sciences, 373*(1752), 20170144.

Beltagy, I., Lo, K., Cohan, A. (2019). Scibert: A pretrained language model for scientific text. In: *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp 3615–3620

Bergelson, E., & Swingley, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences, 109*(9), 3253–3258.

Bergelson, E., & Swingley, D. (2013). The acquisition of abstract words by young infants. *Cognition, 127*(3), 391–397.

Bojanowski, P., Grave, E., Joulin, A., et al. (2017). Enriching word vectors with subword information. *Transactions of the association for computational linguistics, 5*, 135–146.

Bordes, P., Zablocki, E., Soulier, L., et al. (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)*. Association for Computational Linguistics, Hong Kong, China, pp 696–707, 10.18653/v1/D19-1064, https://www.aclweb.org/anthology/D19-1064

Borghi, A. M., Binkofski, F., Castelfranchi, C., et al. (2017). The challenge of abstract concepts. *Psychological Bulletin,143*(3). https://doi.org/10.1037/bul0000089

Borghi, A. M., Barca, L., Binkofski, F., et al. (2019). Words as social tools: Language, sociality and inner grounding in abstract concepts. *Physics of life reviews, 29*, 120–153.

Borghi, A. M., & Binkofski, F. (2014). The Problem of Definition, Springer, New York, NY, pp 1–17. *SpringerBriefs in Psychology*. https://doi.org/10.1007/978-1-4614-9539-0_1

Bracci, S., Ritchie, J. B., Kalfas, I., et al. (2019). The ventral visual pathway represents animal appearance over animacy, unlike human behavior and deep neural networks. *Journal of Neuroscience, 39*(33), 6513–6525.

Breedin, S. D., Saffran, E. M., & Coslett, H. B. (1994). Reversal of the concreteness effect in a patient with semantic dementia. *Cognitive Neuropsychology, 11*(6), 617–660. https://doi.org/10.1080/02643299408251987

Bruni, E., Tran, N. K., & Baroni, M. (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Research, 49*, 1–47.

Brysbaert, M., Stevens, M., Mandera, P., et al. (2016). How many words do we know? practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. *Frontiers in psychology, 7*, 1116.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46*(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5

Bulat L, Clark S, Shutova E (2017) Speaking, seeing, understanding: Correlating semantic models with conceptual representation in the brain. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Copenhagen, Denmark, pp 1081–1091, https://doi.org/10.18653/v1/D17-1113https://aclanthology.org/D17-1113

Cer, D., Diab, M., Agirre, E., et al. (2017). *Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation*. arXiv:1708.00055

Chen, X., Fang, H., Lin, T.Y., et al. (2015). *Microsoft coco captions: Data collection and evaluation server*. arXiv:1504.00325

Cho, K., van Merrienboer, B., Gulcehre, C., et al. (2014a). Learning phrase representations using rnn encoder-decoder for statistical machine translation. In: *Conference on empirical methods in natural language processing* (EMNLP 2014)

Cho, K., Van Merriënboer, B., Bahdanau, D., et al (2014b). *On the properties of neural machine translation: Encoder-decoder approaches*. arXiv:1409.1259

Chrupała G, Kádár Á, Alishahi A (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)*. Association for Computational Linguistics, Beijing, China, pp. 112–118, https://doi.org/10.3115/v1/P15-2019https://www.aclweb.org/anthology/P15-2019

Collell Talleda G, Zhang T, Moens MF (2017) Imagined visual representations as multimodal embeddings. In: *Proceedings of the Thirty-First AAAI conference on artificial intelligence* (AAAI-17), AAAI, pp. 4378–4384

Davis, C. P., Altmann, G. T., & Yee, E. (2020). Situational systematicity: A role for schema in understanding the differences between abstract and concrete concepts. *Cognitive Neuropsychology, 37*(1–2), 142–153.

De Deyne, S., Navarro, D. J., Collell, G., et al. (2021). Visual and Affective Multimodal Models of Word Meaning in Language and Mind. *Cognitive Science,45*(1). https://doi.org/10.1111/cogs.12922

Deng, J., Dong, W., Socher, R., et al. (2009). Imagenet: A large-scale hierarchical image database. In: *2009 IEEE conference on computer vision and pattern recognition*, IEEE, pp. 248–255

Devlin, J., Chang, M.W., Lee, K., et al. (2018). *Bert: Pre-training of deep bidirectional transformers for language understanding*. arXiv:1810.04805

Dolan, W.B., Brockett, C. (2005). Automatically constructing a corpus of sentential paraphrases. In: *Proceedings of the third international workshop on paraphrasing (IWP2005)*

Dove, G. (2018). Language as a disruptive technology: abstract concepts, embodiment and the flexible mind. *Philosophical Trans-*

*actions of the Royal Society B: Biological Sciences, 373*(1752), 20170135.

Dozat, T. (2016). Incorporating nesterov momentum into adam.

Elekes, A., Englhardt, A., Schäler, M., et al. (2018). Resources to examine the quality of word embedding models trained on n-gram data. In: *Proceedings of the 22nd conference on computational natural language learning*, pp. 423–432

Finkelstein, L., Gabrilovich, E., Matias, Y., et al. (2001). Placing search in context: The concept revisited. In: *Proceedings of the 10th international conference on World Wide Web*, pp. 406–414

Gaskell, M. G., & Marslen-Wilson, W. D. (1997). Integrating form and meaning: A distributed model of speech perception. *Language and Cognitive Processes, 12*(5–6), 613–656.

Gatti, D., Marelli, M., Vecchi, T., et al. (2022). Spatial representations without spatial computations. *Psychological Science, 33*(11), 1947–1958.

Gerz, D., Vulić, I., Hill, F., et al. (2016). SimVerb-3500: A large-scale evaluation set of verb similarity. In: *Proceedings of the 2016 conference on empirical methods in natural language processing*. Association for Computational Linguistics, Austin, Texas, pp. 2173–2182, https://doi.org/10.18653/v1/D16-1235https://www.aclweb.org/anthology/D16-1235

Goldstone, R. L. (1995). Effects of Categorization on Color Perception. *Psychological Science,6*(5). https://doi.org/10.1111/j.1467-9280.1995.tb00514.x

Günther, F., Rinaldi, L., & Marelli, M. (2019). Vector-Space Models of Semantic Representation From a Cognitive Perspective: A Discussion of Common Misconceptions. *Perspectives on Psychological Science,14*(6), 1006–1033. https://doi.org/10.1177/1745691619861372. https://journals.sagepub.com/doi/full/10.1177/1745691619861372?casa_token=1ExD6W7c-d8AAAAA%3AMsVbOpTHuyJAeDkB-zrl1okbzKD653dG9rfDxVSDWD8_sx5C8UEux73TV29m3Ep4qaix2Onvck

Günther, F., Petilli, M. A., Vergallito, A., et al. (2022). Images of the unseen: extrapolating visual representations for abstract and concrete words in a data-driven computational model. *Psychological Research*. https://doi.org/10.1007/s00426-020-01429-7

Halawi, G., Dror, G., Gabrilovich, E., et al. (2012). Large-scale learning of word relatedness with constraints. In: *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1406–1414

Harris, Z. S. (1954). Distributional Structure. *WORD, 10*(2–3), 1–2. https://doi.org/10.1080/00437956.1954.11659520

Hasegawa, M., Kobayashi, T., Hayashi, Y. (2017). Incorporating visual features into word embeddings: A bimodal autoencoder-based approach. In: *IWCS 2017 — 12th international conference on computational semantics — short papers*,https://www.aclweb.org/anthology/W17-6912

Heitmeier, M., Chuang, Y. Y., & Baayen, R. H. (2021). Modeling morphology with linear discriminative learning: considerations and design choices. *Frontiers in Psychology*, https://www.frontiersin.org/articles/10.3389

Heitmeier, M., Chuang, Y.Y., Baayen, H. (2023). How trial-to-trial learning shapes mappings in the mental lexicon: Modelling lexical decision with linear discriminative learning. under review. arXiv:2207.00430

Hill, F., Korhonen, A. (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)*, pp. 255–265

Hill, F., Reichart, R., & Korhonen, A. (2015). Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics, 41*(4), 665–695.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation, 9*(8), 1735–1780.

Hoffman, D. (2019). The case against reality: Why evolution hid the truth from our eyes. *WW Norton & Company*.

Hoffman, P., McClelland, J. L., & Lambon Ralph, M. A. (2018). Concepts, control, and context: A connectionist account of normal and disordered semantic cognition. *Psychological Review, 125*(3), 293.

Hollenstein, N., de la Torre, A., Langer, N., et al. (2019). CogniVal: A Framework for Cognitive Word Embedding Evaluation. In: *Proceedings of the 23rd conference on computational natural language learning (CoNLL)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 10.18653/v1/K19-1050

Howell, S. R., Jankowicz, D., & Becker, S. (2005). A model of grounded language acquisition: Sensorimotor features improve lexical and grammatical learning. *Journal of Memory and Language, 53*(2), 258–276.

Husserl, E. (1913). Ideen zu einer reinen Phänomenologie und phänomenologischen Philosophie. Felix Meiner Verlag (2009).

Iki, T., Aizawa, A. (2021). Effect of visual extensions on natural language understanding in vision-and-language models. In: *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 2189–2196

Ioffe, S., Szegedy, C. (2015). Batch normalization: Accelerating deep networks training by reducing internal covariate shift. In: *International conference on machine learning, PMLR*, pp. 448–456

Iverson, J. M., & Goldin-Meadow, S. (2005). Gesture paves the way for language development. *Psychological Science, 16*(5), 367–371.

Johns, B. T., & Jones, M. N. (2012). Perceptual Inference Through Global Lexical Similarity. *Topics in Cognitive Science,4*(1). https://doi.org/10.1111/j.1756-8765.2011.01176.x

Johns, B. T., & Jones, M. N. (2022). Content matters: Measures of contextual diversity must consider semantic content. *Journal of Memory and Language, 123*(104), 313.

Jones, M. N., & Mewhort, D. J. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review, 114*(1), 1.

Kant, I., Guyer, P., & Wood, A. W. (1999). *Critique of pure reason*. Cambridge University Press.

Kiela, D., Bottou, L. (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)*. Association for Computational Linguistics, Doha, Qatar, pp 36–45, 10.3115/v1/D14-1005 https://www.aclweb.org/anthology/D14-1005

Kiela, D., Clark, S. (2015). Multi-and cross-modal semantics beyond vision: Grounding in auditory perception. In: *Proceedings of the 2015 conference on empirical methods in natural language processing*, pp. 2461–2470

Kiela D, Bulat L, Clark S (2015) Grounding semantics in olfactory perception. In: *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (Volume 2: Short Papers)*, pp. 231–236

Kiela D, Conneau A, Jabri A, et al (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)*. Association for Computational Linguistics, New Orleans, Louisiana, pp 408–418, 10.18653/v1/N18-1038 https://www.aclweb.org/anthology/N18-1038

Kiros J, Chan W, Hinton G (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). Association for Computational Linguistics, Melbourne, Australia, pp 922–933. https://doi.org/10.18653/v1/P18-1085https://www.aclweb.org/anthology/P18-1085

Lake, B.M., Murphy, G.L. (2021). Word meaning in minds and machines. *Psychological review*

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by, vol 111.* Chicago London.

Lakoff, G. (1987). Women, Fire, and Dangerous Things. *University of Chicago Press*. https://doi.org/10.7208/chicago/9780226471013.001.0001

Lan, Z., Chen, M., Goodman, S., et al. (2019). Albert: A lite bert for self-supervised learning of language representations. In: *International conference on learning representations*

Landauer, T. K. (1999). Latent Semantic Analysis (LSA), a disembodied learning machine, acquires human word meaning vicariously from language alone. *Behavioral and Brain Sciences, 22*(4). https://doi.org/10.1017/S0140525X99382145

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*(2). https://doi.org/10.1037/0033-295X.104.2.211

Langacker, R. W. (1999). A view from cognitive linguistics. *Behavioral and Brain Sciences, 22*(4). https://doi.org/10.1017/S0140525X99392141

Langacker, R. W. (1987). *Foundations of cognitive grammar: Theoretical prerequisites,* (Vol. 1). Stanford University Press.

Lazaridou, A., Marelli, M., & Baroni, M. (2017). Multimodal Word Meaning Induction From Minimal Exposure to Natural Text. *Cognitive Science, 41*. https://doi.org/10.1111/cogs.12481

Lazaridou A, Chrupała G, Fernández R, et al (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*. Association for Computational Linguistics, Stroudsburg, PA, USA, https://doi.org/10.18653/v1/N16-1043

Lee, J., Yoon, W., Kim, S., et al. (2020). Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics, 36*(4), 1234–1240.

Levesque, H., Davis, E., Morgenstern, L. (2012). The winograd schema challenge. In: *Thirteenth international conference on the principles of knowledge representation and reasoning*

Lin, T.Y., Maire, M., Belongie, S., et al. (2014) Microsoft coco: Common objects in context. In: *European conference on computer vision*, Springer, pp. 740–755

Loshchilov, I., Hutter, F. (2017). *Decoupled weight decay regularization*. arXiv:1711.05101

Louwerse, M. M. (2011). Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science, 3*(2), 273–302.

Louwerse, M., & Connell, L. (2011). A Taste of Words: Linguistic Context and Perceptual Simulation Predict the Modality of Words. *Cognitive Science, 35*(2), 381–398. https://doi.org/10.1111/j.1551-6709.2010.01157.x

Louwerse, M. M., & Zwaan, R. A. (2009). Language Encodes Geographical Information. *Cognitive Science, 33*(1), 51–73. https://doi.org/10.1111/j.1551-6709.2008.01003.x

Lüddecke, T., Agostini, A., Fauth, M., et al. (2019). Distributional semantics of objects in visual scenes in comparison to text. *Artificial Intelligence, 274,*. https://doi.org/10.1016/j.artint.2018.12.009

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers, 28*(2). https://doi.org/10.3758/BF03204766

Luong T, Socher R, Manning C (2013) Better word representations with recursive neural networks for morphology. In: *Proceedings of the seventeenth conference on computational natural language learning*. Association for Computational Linguistics, Sofia, Bulgaria, pp. 104–113. https://www.aclweb.org/anthology/W13-3512

Magnuson, J. S., You, H., Luthra, S., et al. (2020). Earshot: A minimal neural network model of incremental human speech recognition. *Cognitive science, 44*(4), e12823.

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language, 92*, 57–78.

Mannering, W. M., & Jones, M. N. (2021). Catastrophic interference in predictive neural network models of distributional semantics. *Computational Brain & Behavior, 4*(1), 18–33.

Marelli, M., & Amenta, S. (2018). A database of orthography-semantics consistency (osc) estimates for 15,017 english words. *Behavior Research Methods, 50*, 1482–1495.

Martin, A. (2007). The Representation of Object Concepts in the Brain. *Annual Review of Psychology, 58*(1), 25–45. https://doi.org/10.1146/annurev.psych.57.102904.190143

Maruish, M.E., Moses, J.A. (2013). *Clinical neuropsychology: Theoretical foundations for practitioners*. Psychology Press

McRae, K., Cree, G. S., Seidenberg, M. S., et al. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods, 37*(4). https://doi.org/10.3758/BF03192726

Merleau-Ponty, M., Landes, D., Carman, T., et al. (2013). *Phenomenology of perception*. Routledge.

Mestres-Missé, A., Münte, T. F., & Rodriguez-Fornells, A. (2014). Mapping concrete and abstract meanings to new words using verbal contexts. *Second Language Research, 30*(2), 191–223. https://doi.org/10.1177/0267658313512668

Mikolov, T., Chen, K., Corrado, G., et al. (2013). Efficient Estimation of Word Representations in Vector Space. *International conference on learning representations*

mir. (2008). *Acquisition and representation of grammatical categories: Grammatical gender in a connectionist network*

Mkrtychian N, Blagovechtchenski E, Kurmakaeva D, et al (2019) Concrete vs. Abstract Semantics: From Mental Representations to Functional Brain Mapping. Frontiers in Human Neuroscience 13(August):267. 10.3389/fnhum.2019.00267, https://www.frontiersin.org/article/10.3389/fnhum.2019.00267

Montefinese, M. (2019). Semantic representation of abstract and concrete words: A minireview of neural evidence. *Journal of Neurophysiology, 121*(5), 1585–1587. https://doi.org/10.1152/jn.00065.2019, https://journals.physiology.org/doi/full/10.1152/jn.00065.2019

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

Paivio, A. (1971). Imagery and verbal processes. new york, ny: Holt, rinheart & winston. paivio, a. 1986. *Mental representation: A dual-coding approach*

Paivio, A. (1965). Abstractness, imagery, and meaningfulness in paired-associate learning. *Journal of Verbal Learning and Verbal Behavior, 4*(1), 32–38. https://doi.org/10.1016/S0022-5371(65)80064-0

Park, J., Myaeng, Sh. (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)*. Asian Federation of Natural Language Processing, Taipei, Taiwan, pp 214–223, https://www.aclweb.org/anthology/I17-1022

Pearson, K. (1901). Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science, 2*(11), 559–572.

Pennington, J., Socher, R., Manning, C. (2014). Glove: Global Vectors for Word Representation. In: *Proceedings of the 2014*

*conference on empirical methods in natural language processing (EMNLP)*. Association for Computational Linguistics, Stroudsburg, PA, USA, https://doi.org/10.3115/v1/D14-1162

Peterson, J.C., Abbott, J.T., Griffiths, T.L. (2017). Adapting Deep Network Features to Capture Psychological Representations: An Abridged Report. In: *Proceedings of the twenty-sixth international joint conference on artificial intelligence*. International Joint Conferences on Artificial Intelligence Organization, California, https://doi.org/10.24963/ijcai.2017/697

Petilli, M. A., Günther, F., Vergallito, A., et al. (2021). Data-driven computational models reveal perceptual simulation in word processing. *Journal of Memory and Language,117*,. https://doi.org/10.1016/j.jml.2020.104194

Pezzelle, S., Takmaz, E., Fernández, R., et al. (2021). Word representation learning in multimodal pre-trained transformers: An intrinsic evaluation. *Transactions of the association of computational linguistics* 9

Pierrejean, B., Tanguy, L. (2019). Investigating the stability of concrete nouns in word embeddings. In: *Proceedings of the 13th international conference on computational semantics-short papers*, pp. 65–70

Rajpurkar, P., Zhang, J., Lopyrev, K., et al. (2016). *Squad: 100,000+ questions for machine comprehension of text*. arXiv:1606.05250

Reif, E., Yuan, A., Wattenberg, M., et al. (2019). Visualizing and measuring the geometry of bert. *Advances in neural information processing systems* 32

Rotaru, A. S., & Vigliocco, G. (2020). Constructing semantic models from words, images, and emojis. *Cognitive science, 44*(4), e12830.

Schwanenflugel, P. J., Akin, C., & Luh, W. M. (1992). Context availability and the recall of abstract and concrete words. *Memory & Cognition, 20*(1), 96–104. https://doi.org/10.3758/BF03208259

Schwanenflugel, P. J., & Shoben, E. J. (1983). Differential context effects in the comprehension of abstract and concrete verbal materials. *Journal of Experimental Psychology: Learning, memory, and cognition, 9*(1), 82–102. https://doi.org/10.1037/0278-7393.9.1.82

Schwanenflugel, P. J., & Stowe, R. W. (1989). Context Availability and the Processing of Abstract and Concrete Words in Sentences. *Reading Research Quarterly, 24*(1), 114. https://doi.org/10.2307/748013

Shahmohammadi H, Lensch HPA, Baayen RH (2021) Learning zero-shot multifaceted visually grounded word embeddings via multi-task training. In: *Proceedings of the 25th conference on computational natural language learning*. Association for Computational Linguistics, Online, pp. 158–170, https://doi.org/10.18653/v1/2021.conll-1.12, https://aclanthology.org/2021.conll-1.12

Silberer C, Lapata M (2014) Learning grounded meaning representations with autoencoders. In: *Proceedings of the 52nd annual meeting of the association for computational linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Baltimore, Maryland, pp 721–732, https://doi.org/10.3115/v1/P14-1068, https://www.aclweb.org/anthology/P14-1068

Sileo, D. (2021). Visual grounding *strategies for text-only natural language processing*. arXiv:2103.13942

Simmons, W. K., Martin, A., & Barsalou, L. W. (2005). Pictures of Appetizing Foods Activate Gustatory Cortices for Taste and Reward. *Cerebral Cortex, 15*(10), 1602–1608. https://doi.org/10.1093/cercor/bhi038

Smith, L., & Gasser, M. (2005). The development of embodied cognition: Six lessons from babies. *Artificial Life, 11*(1–2), 13–29. https://doi.org/10.1162/1064546053278973

Socher, R., Perelygin, A., Wu, J., et al. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In: *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642

Solomon, K. O., & Barsalou, L. W. (2001). Representing Properties Locally. *Cognitive Psychology, 43*(2), 129–169. https://doi.org/10.1006/cogp.2001.0754

Solomon, K. O., & Barsalou, L. W. (2004). Perceptual simulation in property verification. *Memory & Cognition, 32*(2), 244–259. https://doi.org/10.3758/BF03196856

Srivastava, N., Hinton, G., Krizhevsky, A., et al. (2014). Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research, 15*(1), 1929–1958.

Strubell E, Ganesh A, McCallum A (2019) Energy and policy considerations for deep learning in NLP. In: *Proceedings of the 57th annual meeting of the association for computational linguistics*. Association for Computational Linguistics, Florence, Italy, pp. 3645–3650, https://doi.org/10.18653/v1/P19-1355https://aclanthology.org/P19-1355

Student. (1908). The probable error of a mean. *Biometrika* pp. 1–25

Su, W., Zhu, X., Cao, Y., et al. (2019). *Vl-bert: Pre-training of generic visual-linguistic representations*. arXiv:1908.08530

Szegedy, C., Vanhoucke, V., Ioffe, S., et al. (2016). Rethinking the inception architecture for computer vision. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826

Tan, H., Bansal, M. (2020). Vokenization: Improving language understanding with contextualized, visual-grounded supervision. In: *Proceedings of the 2020 Conference on empirical methods in natural language processing (EMNLP)*. Association for Computational Linguistics, Online, pp 2066–2080, https://aclanthology.org/2020.emnlp-main.162

Tyler, L. K., Moss, H. E., & Jennings, F. (1995). Abstract word deficits in aphasia: Evidence from semantic priming. *Neuropsychology, 9*(3), 354.

Utsumi, A. (2022). A test of indirect grounding of abstract concepts using multimodal distributional semantics. *Frontiers in Psychology* 13

VandenBos, G.R. (2015). *APA Dictionary of Psychology*, 2nd edn. American Psychological Association, Washington, DC, http://www.jstor.org/stable/j.ctv1chrw2d

Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. In: *Advances in neural information processing systems*, pp 5998–6008

Von Ahn, L. (2006). Games with a purpose. *Computer, 39*(6), 92–94.

Wang, A., Singh, A., Michael, J., et al. (2018). *Glue: A multi-task benchmark and analysis platform for natural language understanding*. arXiv:1804.07461

Wang, J., Song, Y., Leung, T., et al. (2014). Learning fine-grained image similarity with deep ranking. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 1386–1393

Wang, B., Wang, A., Chen, F., et al. (2019). Evaluating word embedding models: Methods and experimental results. *APSIPA transactions on signal and information processing* 8

Warrington, E. K. (1975). The selective impairment of semantic memory. *The Quarterly journal of experimental psychology, 27*(4), 635–657. https://doi.org/10.1080/14640747508400525

Warstadt, A., Singh, A., & Bowman, S. R. (2019). Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics, 7*, 625–641.

Westbury, C., & Hollis, G. (2019). Wriggly, squiffy, lummox, and boobs: What makes some words funny? *Journal of Experimental Psychology: General,148*(1). https://doi.org/10.1037/xge0000467

Westbury, C. (2014). You Can't rink a Word: Lexical and Individual Emotionality Affect Subjective Familiarity Judgments. *Journal of Psycholinguistic Research, 43*(5), 1–2. https://doi.org/10.1007/s10936-013-9266-2

Wiemer-Hastings K, Krug J, Xu X (2001) Imagery, Context Availabilty, Contextual Constraint and Abstractness. In: *Proceedings of the*

*annual meeting of the cognitive science society*, vol 23. Lawrence Erlbaum, Mahwah, NJ, pp. 1134–1139

Williams, A., Nangia, N., Bowman, S.R. (2017). *A broad-coverage challenge corpus for sentence understanding through inference*. arXiv:1704.05426

Wolf, T., Debut, L., Sanh, V., et al. (2019). *Huggingface's transformers: State-of-the-art natural language processing*. arXiv:1910.03771

Xu, B., Wang, N., Chen, T., et al. (2015). *Empirical evaluation of rectified activations in convolutional network*. arXiv:1505.00853

Yun T, Sun C, Pavlick E (2021) Does vision-and-language pretraining improve lexical grounding? In: *Findings of the association for computational linguistics: EMNLP 2021*. Association for Computational Linguistics, Punta Cana, Dominican Republic, pp. 4357–4366, https://doi.org/10.18653/v1/2021.findings-emnlp.370https://aclanthology.org/2021.findings-emnlp.370

Zeno, S., Ivens, S.H., Millard, R.T., et al. (1995). The educator's word frequency guide. *Touchstone applied science associates*

Zhang, R., Isola, P., Efros, A.A., et al. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In: *CVPR*

Zhu Y, Kiros R, Zemel R, et al (2015) Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In: *The IEEE international conference on computer vision (ICCV)*

Zwaan RA, Madden CJ (2005) Embodied Sentence Comprehension. In: *Grounding cognition*. Cambridge University Press, https://doi.org/10.1017/CBO9780511499968.010