Explainable Semantic Space by Grounding Language to Vision with Cross-Modal Contrastive Learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

In natural language processing, most models try to learn semantic representations merely from texts. The learned representations encode the "distributional semantics" but fail to connect to any knowledge about the physical world. In contrast, humans learn language by grounding concepts in perception and action and the brain encodes "grounded semantics" for cognition. Inspired by this notion and recent work in vision-language learning, we design a two-stream model for grounding language learning in vision. The model includes a VGG-based visual stream and a Bert-based language stream. The two streams merge into a joint representational space. Through cross-modal contrastive learning, the model first learns to align visual and language representations with the MS COCO dataset. The model further learns to retrieve visual objects with language queries through a cross-modal attention module and to infer the visual relations between the retrieved objects through a bilinear operator with the Visual Genome dataset. After training, the model's language stream is a stand-alone language model capable of embedding concepts in a visually grounded semantic space. This semantic space manifests principal dimensions explainable with human intuition and neurobiological knowledge. Word embeddings in this semantic space are predictive of human-defined norms of semantic features and are segregated into perceptually distinctive clusters. Furthermore, the visually grounded language model also enables compositional language understanding based on visual knowledge and multimodal image search with queries based on images, texts, or their combinations.

1 Introduction

Humans take much longer time to name a colored word when the color and the word mismatch (e.g., "*red*" shown in green) than when they match (e.g., "*red*" shown in red) [1]. This effect is an example of rich psychological evidence suggesting that humans learn language by grounding meanings to knowledge about the world [2, 3]. In contrast, most models in natural language processing (NLP) [4–7] encode "distributional semantics" [8] learned from texts only. It is arguably like putting machines in a thought experiment for the "Chinese Room Argument" [9]. Imagine that you have to learn Chinese from scratch as your first language. All that you have is a Chinese-to-Chinese dictionary. You might be able to relate a word to other words based on textual distributions. It is, however, impossible to learn word meanings without any additional explanation in reference to the physical world [10].

Instead, a language model may learn concepts from texts paired with sensory data, such as images. This notion of joint vision-language learning has been explored for specific goals, e.g., image captioning [11], visual question answering [12], etc. In line with these studies, we train a language model and a vision model jointly to match images and texts in terms of representations of images,

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

natural language, visual objects and their relationships. In this way, the language model learns a new semantic space grounded in vision. In this space, semantic embeddings are found to be organized and clustered by visual attributes, predictive of human-defined norms of semantic features, useful for compositional language understanding and cross-modal image search. In addition, this visually grounded language model is potentially useful to understand the computational basis of grounded cognition [13, 14].

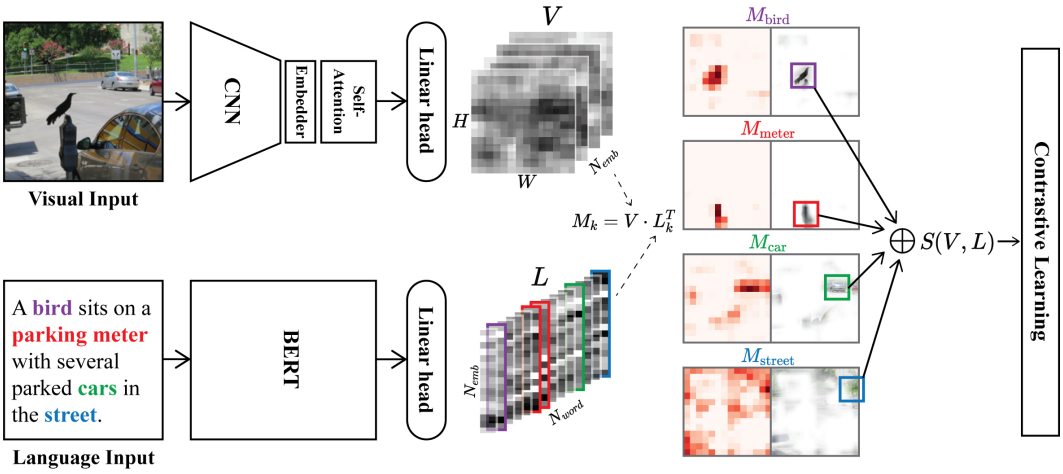


Figure 1: Visual grounding model for natural language (see Section 3.1). The visual and language streams take an image and its caption as input, respectively. The inner-product between the visual feature maps and the contextual word embeddings forms the 3D match-map that highlights the matching between visual and language content. The similarity score calculated from the match-map (see Eq. 1) is used to evaluate the cross-modal contrastive loss.

2 Background and Related Work

2.1 Distributional vs. grounding hypothesis

In the *distributional hypothesis* [15], words that occur in similar contexts carry similar meanings. This hypothesis has motivated influential machine learning models to learn word embeddings from large text corpora [4, 16]. However, the learned word embeddings are not straightforward to interpret [8, 17]. Alternatively, the *symbol grounding hypothesis* suggests that a word is connected to its meaning by relating to its referent in the physical world [10, 18]. In line with this hypothesis, earlier studies demonstrate that visual features or contexts can enhance language learning [19–23].

2.2 Vision-language learning

A strategy for grounded language learning extends the distributional hypothesis from *unimodal* contexts to *multimodal* contexts [24–26]. For example, a word that describes an object may occur in the context of an image pertaining to that object's specific visual features, e.g., the color, shape, texture, and environment. When an image and its caption are used together for training a language model, the visual features may potentially reshape the semantic representations to be more explainable [27]. Inspired by the success of the attention-based transformer in NLP [6, 28], recent work in vision-language learning also applies the attention mechanism to paired images and texts and show strong performance in cross-modal tasks [29–32] .

Contrastive learning [33, 34] is increasingly applied to not only unimodal data [35] but also multi-modal data [36, 37]. However, it remains underexplored whether cross-modal representations learned with contrastive learning can support higher-level tasks, e.g., visual question answering [12], visual reasoning [38], scene graph generation [39]. Such tasks involve abstract reasoning about the relations between entities (e.g., visual objects). Prior work approaches relational inference with multi-layer perceptrons [40, 41] or graph neural networks [42–44]. A more compelling idea [45] is to model

entities as vectors in a continuous space and to model their relations as arithmetic operators (linear [46–48] or bilinear [49, 50]) applied to the vector representations of those entities.

2.3 Relation to prior work

To apply visual grounding [21] to the state-of-the-art language learning models [6], we first build a two-stream model to jointly learn visual and language representation from image-caption pairs. We then finetune the learned model by adding a cross-modal attention layer [29, 30] and bilinear operators [49], which are used to extract representations of the relations between visual objects. Both stages of learning utilize contrastive loss [36, 37]. The visual and language streams are separable and computable as stand-alone systems after training, unlike some vision-language models that require both visual and textual input in order to be usable [29, 30]. Closely related to our work, Harwath and colleagues match visual objects to spoken words using triplet loss [36]. Our work takes one step further by utilizing contrastive learning to learn relations between visual objects. In another closely related work, Jia and colleagues use cross-modal contrastive learning to pretrain a vision model with noisy text supervision via leveraging a massive image-text dataset [37]. Here, we focus on the language model learned with visual data and characterize it as a stand-alone system as well as a system for cross-modal tasks. We perform a systematic evaluation of the semantic space grounded in vision vs. the ungrounded semantic space learned from texts only. Following the similar goal in prior work [51], we aim to assess how visual grounding affects the distribution of textual representations.

3 Approach

3.1 Visual grounding of natural language

To build a computational model for learning visually grounded language representations, we develop a model (Fig. 1) that combines a stand-alone visual stream and a stand-alone language stream. The visual stream is based on VGG16 [52] with an additional linear transformation as an embedder to match the feature dimension of the language stream and an additional multi-head self-attention layer [53] to enforce global information aggregation and learn long-range dependency. The language stream is based on Bert [6]. Using separate linear transformation heads [35], the output from both the visual stream and the language stream are projected to a common representational space. In this common space, the inner-product between the visual representation $\boldsymbol{V}$ at every location and the language representation $\boldsymbol{L}$ of every word gives rise to a 3D match-map, where each element indicates how a word in the text matches each location in the image (See illustration in Fig. 1). The sum of the maximal match is the similarity score $S(\boldsymbol{V}, \boldsymbol{L})$ between a pair of image and text (Eq. 1; i, j indicate the location in the 2D image feature map $\boldsymbol{V}$, k indicates the k-th word in $\boldsymbol{L}$).

$$\boldsymbol{M}_{i,j,k} = \boldsymbol{V}_{i,j} \cdot \boldsymbol{L}_k^T, \quad S(\boldsymbol{V}, \boldsymbol{L}) = \sum_{k=1}^{K} \max_{i,j} \boldsymbol{M}_{i,j,k} \tag{1}$$

Based on the unimodal normalized temperature-scaled cross-entropy (NT-Xent) loss [35, 37, 54], we can construct the cross-modal contrastive loss using the anchor sample from one modality and the positive sample and negative samples from the other modality. As such, we define and sum two loss functions with the anchor sample from either images or texts and positive/negative samples from either texts or images, respectively.

$$\text{Loss}_l = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp(S(\boldsymbol{V}_i, \boldsymbol{L}_i)/\tau)}{\sum_{j=1}^{B} \exp(S(\boldsymbol{V}_i, \boldsymbol{L}_j)/\tau)}, \quad \text{Loss}_v = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{\exp(S(\boldsymbol{V}_i, \boldsymbol{L}_i)/\tau)}{\sum_{j=1}^{B} \exp(S(\boldsymbol{V}_j, \boldsymbol{L}_i)/\tau)} \tag{2}$$

For Loss_l in Eq. 2, the anchor sample $\boldsymbol{V}_i$ is an input image and the positive sample $\boldsymbol{L}_i$ is the corresponding image caption, whereas the negative samples $\boldsymbol{L}_j$ are unmatched textual descriptions included in the same batch (B is the batch size). Similarly, Loss_v in Eq. 2 is defined to contrast the positive and negative image samples against an anchor textual sample.

3.2 Visual grounding of object relations

We further finetune the model for visual relation prediction, as illustrated in Fig. 2. In this stage, we remove the linear transformation heads in Fig. 1 and add a multi-head cross-modal attention module [29, 30]. The attention module uses a query based on the embedding of an object word from the language stream (Query$_L$) and uses keys (Key$_V$) and values (Value$_V$) from every location of the feature map output from the visual stream. The attention score is calculated as the inner-product of Query$_L$ and Key$_V$ followed by softmax. The attention-weighted sum of Value$_V$ is concatenated across 8 attention heads to generate a visually grounded object representation.

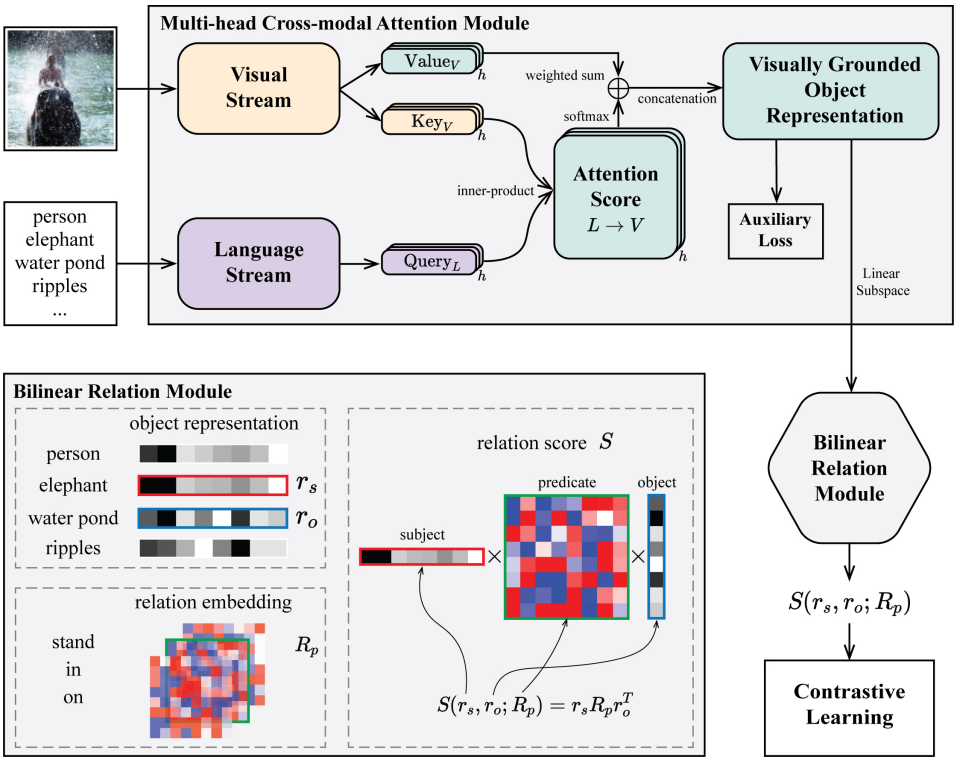


Figure 2: Model architecture for visual grounding of object relation. The language stream uses an object description as input (e.g., large black elephant; we only show the object name "elephant" in this illustration for simplicity). The multi-head cross-attention module outputs a set of visually grounded object representations (See detailed methods in Appendix A.3). The bilinear relation module (bottom left) further generates a relation score given representations of a (**s**ubject,**p**redicate,**o**bject) triplet (e.g. `(elephant,in,water pond)`) for contrastive learning.

Applied to the grounded object representation is a bilinear relation module for predicting the visual relation between two objects (linguistically a subject and an object). For both the subject and the object, their grounded representations are linearly transformed to a subspace $D = \mathbb{R}^d$ (default $d = 32$), denoted as r_s and r_o, respectively. A predicate p is represented as a learnable bilinear operator $F_p :< D, D >\rightarrow \mathbb{R}$, which represents the relation embedding R_p. Applying this bilinear operator to the subject vs. object representations measures their relation score S specific to the given predicate [49] expressed as (Eq. 3).

$$S(r_s, r_o; R_p) = F_p(r_s, r_o) = r_s R_p r_o^T \tag{3}$$

For visual relation prediction, we also use contrastive learning with two loss functions by taking either relation embedding or subject/object representations as positive/negative samples.

$$\text{Loss}_{\text{rel}} = -\frac{1}{|\mathcal{B}|} \sum_{(r_s, r_o; R_p) \in \mathcal{B}} \log \frac{\exp(S(r_s, r_o; R_p)/\tau)}{\sum_{k \in \mathcal{K}_{\text{rel}}} \exp\big(S(r_s, r_o; R_p^k)/\tau\big)} \tag{4}$$

$$\text{Loss}_{\text{obj}} = -\frac{1}{|\mathcal{B}|} \sum_{(r_s, r_o; R_p) \in \mathcal{B}} \log \frac{\exp(S(r_s, r_o; R_p)/\tau)}{\sum_{k \in \mathcal{K}_{\text{obj}}} \exp(S(r_s^k, r_o^k; R_p)/\tau)} \tag{5}$$

In Loss_{rel}, $\mathcal{K}_{\text{rel}}$ is the set that contains all relations available. The anchor sample is a pair of subject and object in an image. The positive sample is the embedding of the ground truth relation. The negative samples are the embeddings of all other relations. In Loss_{obj}, the anchor sample is a given relation. The positive sample is a subject-object pair that holds this relation. The negative samples are other subject-object pairs in a different relation. For both loss functions, the positive and negative samples are drawn from the same batch $\mathcal{B}$.

In addition, we also add a classification head (two fully connected layers with ReLU in between) and apply it to the grounded object representation. We use object classification as an auxiliary objective (with a cross-entropy loss) to constrain the grounded object representation to be separable across objects for classification.

3.3 Training and Testing

We train the model in three stages to progressively refine the model with increasingly demanding tasks. In the first stage, we pretrain the visual and language streams separately as image and text encoders. The language stream is the pretrained Bert[1] used as the baseline model for subsequent experiments. The visual stream is pretrained for object classification with ImageNet [55]. Relative to the baseline CNN, the inclusion of self-attention improves the top-1 classification accuracy from 71.6% to 74.3% on the ImageNet validation dataset. The attention module also renders the classification more robust when the input image is partially occluded (See details in Appendix A.1).

In the second stage, we refine the pretrained language and visual streams by matching texts to images, as illustrated in Fig. 1 on the MS COCO dataset [11]. While freezing other layers, we refine the self-attention layer in the visual stream and the top k layers in Bert (by default $k = 8$). Training with contrastive learning is based on the MS COCO dataset. As five captions are available for each image, we randomly sample one caption per image in each iteration. The results suggest earlier grounding (larger k) tends to support better image-text retrieval performance (see details in Appendix A.2).

In the third stage, we further finetune the model for visual relation prediction as illustrated in Fig. 2. We refine the visual self-attention layer and the higher l layers in Bert (by default $l = 2$) based on the Visual Genome dataset [39] after cleaning the dataset to include 114 relations and 55 object classes in order to alleviate imbalanced data across different classes or relations (Appendix A.3). The training does not use any image annotation (e.g., bounding box) and is independent of other models (e.g., object detection). Instead, we use cross-attention to retrieve visual objects from raw images given textual queries and learn the representations of the retrieved objects and their relations altogether. After training, the model predicts the object class with 97.71% top-1 accuracy and predicts the relation label with 64.26% top-1 accuracy (See details and examples in Appendix A.3).

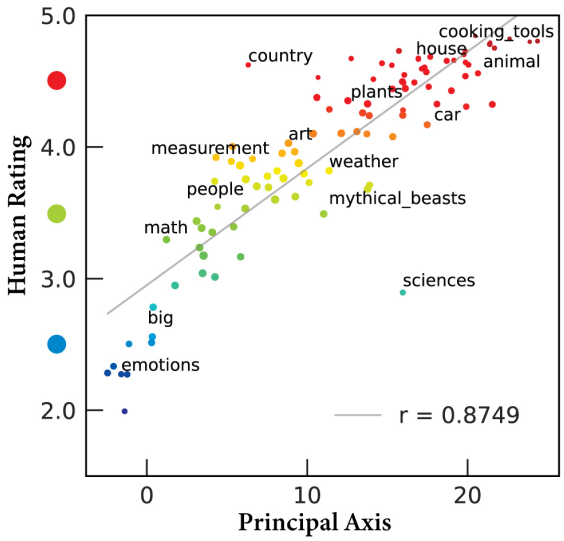
4 Experiments

4.1 Principal components of grounded semantic representations

To evaluate the visually grounded semantic space, we use the language stream as a stand-alone model to extract the output representations of commonly used English words in the SemCat dataset ($9,197$ words; 100 word categories) [56] (details about how word representations are extracted from the language stream are explained in Appendix B). We apply the principal component analysis to the representations of all the words studied here and examine the top components as the principal dimensions of the grounded semantic space.

Interestingly, the first principal dimension is readily interpretable as an abstract-to-concrete axis (Fig. 3). For example, words with the highest values in this axis are *ostrich, seagull, albatross, blender, pelican, broccoli, parakeet, lettuce, sailboat, vegetables*, whereas words with the lowest values are *displeasure, liking, to, outgoing, present, experienced, profitable, faithful, meaningful, multitude*. The representations of words along this axis is significantly correlated with human rating

[1] `bert-base-uncased`: https://huggingface.co/transformers/pretrained_models.html



Examples:

word	category	principal axis	human rating
oven	cooking tool	30.86	4.97
zebra	animal	24.74	4.86
car	car	22.06	4.89
furniture	house	20.83	4.89
defense	country	4.65	4.19
chemistry	sciences	18.67	3.64
wood	plants	16.22	4.85
cartoon	art	11.84	4.33
humid	weather	11.58	3.48
angel	mythical beasts	7.22	3.82
lover	people	6.01	3.68
thousand	math	3.65	3.07
huge	big	0.17	3.54
cheerful	emotions	- 2.54	2.34

Figure 3: The first principal component in the grounded semantic space captures the concrete-abstract axis of semantics. Left: Each dot represents a word category with the color indicative of the averaged human-rated concreteness (the y axis) and the size proportional to the standard deviation. The x axis indicates the value projected onto the first principal axis. Right: Example words in labeled categories.

of their concreteness (ranging from 1 to 5) from prior study [57] (Fig. 3). The Pearson correlation coefficient reaches 0.8749 or 0.6948 across word categories or words, respectively (After grounding with object relations: $r = 0.8001$ for categories, $r = 0.6948$ for words). In contrast, the principal axis of the ungrounded semantic space learned from the baseline Bert model is not straightforward to interpret and shows a weak correlation with human ratings of concreteness ($r = 0.3538$ for categories, $r = 0.1040$ for words). Other principal components are also intuitively explainable. For example, PC 2 captures the human vs. non-human axis, PC 3 captures the scene vs. object axis, PC 4 captures the natural vs. artificial axis, PC 5 captures the indoor vs. outdoor axis, PC 6 highlights words related to food (see detailed results about other principal components in Appendix B.1).

4.2 Relation to human-defined norms of semantic features

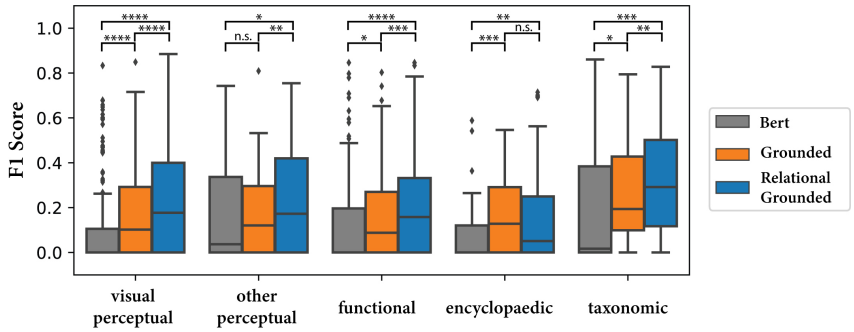


Figure 4: The F1 score of predicting semantic feature norms from word representations before and after visual grounding. Significant level: n.s.: not significant; *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$; ****, $p < 0.0001$.

We further ask whether the visually grounded word embeddings are amenable to binary semantic features defined by humans [23, 58]. We use the concept property norm dataset from the Centre for Speech, Language and the Brain (CSLB) [59]. The dataset includes binary semantic features (e.g., `has_wheels`) labeled for 638 concepts collected from 123 human participants. We keep 390 features

that each contains at least 5 samples. We hypothesize that the grounded word embeddings can be read out through a linear and sparse projection to readily support binary classification attainable by humans. To test this hypothesis, we train a logistic regression model with L1 regularization to predict each binary semantic feature from the grounded word embeddings and repeat this also for ungrounded semantics for comparison. See Appendix B2 for details about this dataset and our evaluation method. Results suggest that the grounded word embeddings are significantly more predictive of visually relevant binary features than ungrounded counterparts obtained by Bert (WilcoxonSigned Rank Test; $p < 0.0001$) (Fig. 4). This difference is less pronounced but still significant for other features related to other perceptual (e.g., `has_flavors`), functional (e.g., `does_cut`), encyclopaedic (e.g., `is_dangerous`), and taxonomic features (e.g., `is_clothing`), especially after visual grounding of object relations.

4.3 Clustering of word representations

After visual grounding, the semantic representations group themselves based on perceptual similarity. We use the SemCat dataset ($9,197$ English words from $N = 100$ categories) [56] and calculate the Silhouette coefficient (between -1 and 1) to measure the degree by which these words are clustered by categories. The distance between word embeddings is measured as the cosine distance (See details in Appendix B3). The Silhouette coefficients across 100 categories are significantly higher for the visually grounded semantics than ungrounded ones (Wilcoxon Signed Rank Test; $p < 0.0001$) (Fig. 5 left). The greatest gain in clustering are noticeable for categories that include concrete concepts (e.g. `car`, `housing`, `mammal`) with defining visual attributes (Fig. 5 right). For some abstract categories related to human emotion (e.g., `happy`), the grounded representations are also better clustered than the ungrounded ones.

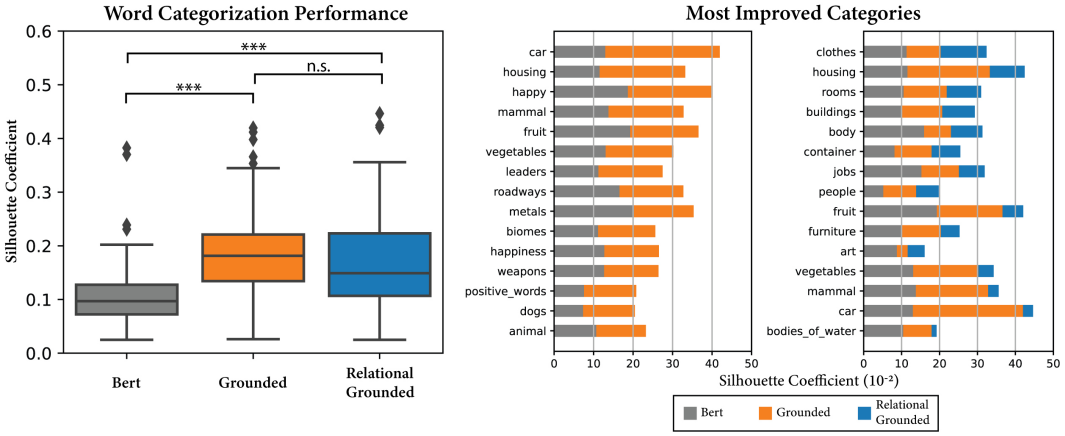


Figure 5: Left: A boxplot showing silhouette coefficients on word representations before and after visual grounding (whiskers=1.5). Right: Top-15 word categories that are better clustered after visual grounding of natural language and object relations.

Some general categories show further fine-grained clusters within themselves in the grounded semantic space. For example, a `boat` is a vehicle without wheels; a `car` is a vehicle that typically has four wheels; an `airplane` is a vehicle with wings. We take these three words (`boat`, `car`, `airplane`) as the prototypes of different vehicles and calculate the cosine similarity between all words in the Semcat dataset and each of these prototype words. Words that are similar to one prototype are notably less similar to other prototypes after visual grounding, setting them apart into clusters typical to `water/land/air` transportation (Fig. 6). List of words in these subcategories and results for other examples (e.g., `dog`, `goose`, `horse` are subcategories of `mammal`) are ellaborated in Appendix B.3.

4.4 Vision based compositional reasoning

A drawback of distributional semantics is the inability to make visually informed compositional reasoning. We know that "zebra is a horse with black and white stripes", because we have seen

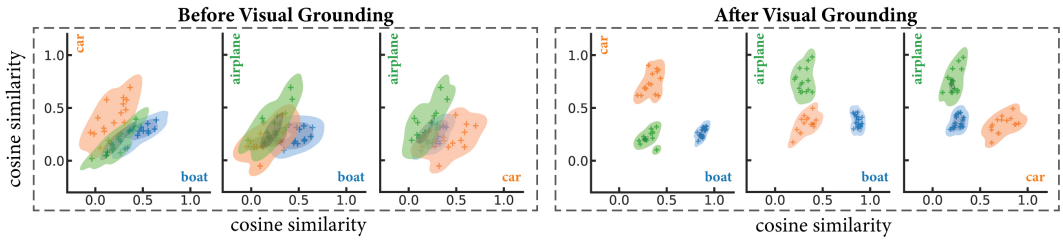


Figure 6: Visualization of words related to vehicle subcategories. Each plot shows a 2D plane expanded by the cosine similarity scores according to a pair of prototype words. Each dot represents a word color-coded by the prototype word (blue:boat; orange:car; green:airplane).

how zebra looks like, whereas an ungrounded language model is never or rarely exposed to such information [10]. We test whether the visually grounded semantics can perform compositional reasoning based on visual knowledge, without being explicit trained to do so. We choose some words (Table 1), for which the meaning can be intuitively inferred from the combination of other words. For example, we use a phrase "striped horse" as a compositional query to search for the matched words ranked in terms of cosine similarity. With the grounded semantic representation, the phrase `striped horse` is highly similar to the word `zebra` (cosine similarity: 0.63), which is ranked as the 8-th among all common English words. Other words similar to `striped horse` all refer to animals within horse familiars (See details in Appendix 5). This is in shape contrast to the ungrounded semantic space, in which it is impossible to relate `striped horse` to `zebra` based on the similarity of their representations (cosine similarity: 0.12; rank: 2,914). See other examples in (Table 1).

Table 1: Examples of vision based conceptual composition. Each row shows the cosine similarity and its ranking in the vocabulary (unique words in the Semcat dataset) between a pair of query phrase and target word. Except (`hot weather`, `summer`), all others are concepts supported by composition of *visual* knowledge in the query phrase.

Query Phrase	**Target Word**	**Similarity (cosine | rank)**					
		Bert		Grounded		Relational	
striped horse	zebra	0.12	2914	0.60	12	**0.63**	**8**
black and white bear	panda	0.13	2478	0.69	**2**	**0.81**	**2**
flying car	plane	0.36	167	**0.66**	**4**	0.61	11
round container	bowl	0.25	489	0.56	8	**0.67**	**2**
red fruit	strawberry	0.39	239	0.75	**3**	**0.85**	**3**
young dog	puppy	0.40	94	0.92	**2**	**0.93**	**2**
iced mountain	glacier	0.44	20	**0.86**	**1**	0.73	5
clear sky	sunny	0.27	631	0.31	184	**0.34**	**61**
hot weather	summer	0.27	903	0.52	14	**0.53**	**6**

4.5 Multimodal image search

The cross-attention module in our model forms a joint representational space to combine both visual and textual input. We explore whether this joint space can be used to support cross-modal tasks, e.g., image search based on image, text, or their combinations [37]. For this task, we add two additional heads. Each head includes two linear layers with ReLU in between followed by average pooling. It is applied to either visual or textual representations in the joint space and results in a single vector representation for an image or a text. While freezing our model described in Section 3.2, we train the two additional heads with contrastive loss to match the average-pooled representations of paired images and texts in terms of their cosine similarity using the MS COCO dataset. To use the model for image search, we apply weighted sum to the normalized representations of a query image and a query text (with weights: $1-\alpha$ and α, where $0 \leq \alpha \leq 1$). We use the multimodal query to search a held-out database[2] for the matched images ranked in terms of cosine similarity.

[2] 41, 600 images from the validation dataset of Open Images Dataset $V6$.

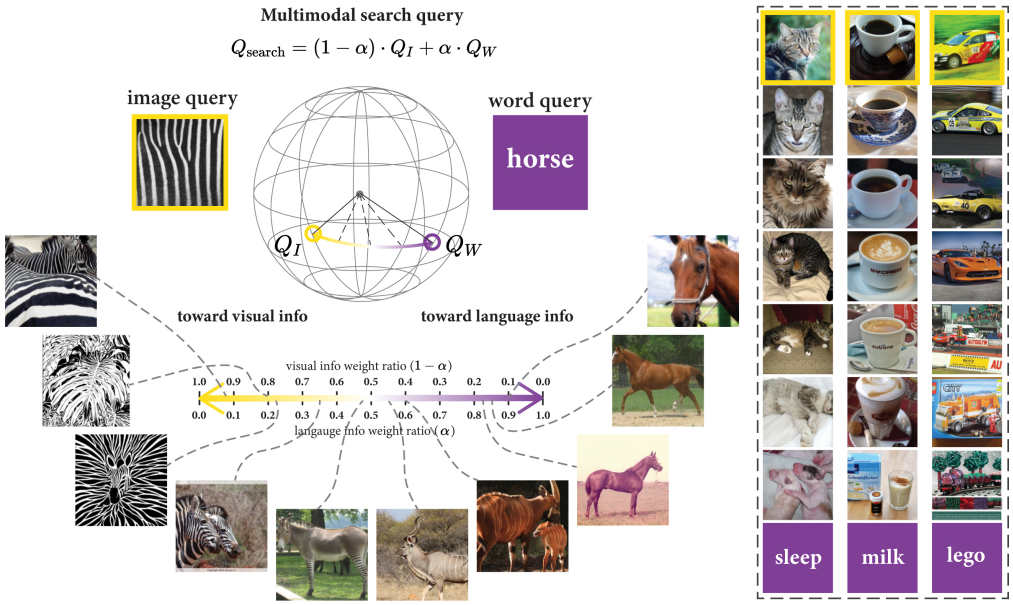


Figure 7: Left: Illustration of multimodal image search with a "zebra" example. The image query Q_I is the L2-normalized vector representation of a zebra's skin pattern. The word query Q_W is the L2-normalized vector representation of word *horse*. As the weight ratio α in multimodal query Q_{search} increases from 0 to 1, the search results show grade change from zebra-like patterns, to a real zebra, and to a horse image. Right: Example multimodal image search results on image "cat" and word *sleep*, image "coffee" and word *milk*, image "car" and word *lego*. Increasing α from 0 to 1 gives the image search results from top to bottom.

As α controls the weighting between the textual and visual queries, we test how the image search returns different results as α increases from 0 (image only) to 1 (text only). For example, when we combine a word (*horse*) and an image (a stripped pattern) into a query, the search finds images similar to the zebra's skin pattern when α is close to 0, or finds images of typical horses when α is close to 1, but not necessarily a zebra for either case, except when α is somewhere close to 0.5 (Fig. 7, left). This observation is generalizable to other examples. See similarly grade changes in (Fig. 7, right).

5 Discussion

Grounding language learning in vision renders semantic representations more explainable by human intuition. The grounded semantic space has its principal dimension encoding the concrete-to-abstract variation, consistent with human ratings and neurobiological knowledge. The grounded semantic representations are better clustered by finer categories and capable of compositional reasoning (e.g., `zebra = striped horse`). In addition, our model can represent and combine visual and textual information in a joint and continuous representational space to support multimodal image search.

The visually grounded language model may be usable as a computational model for studying the grounded cognition - a theory in cognitive science [13, 14]. Ungrounded linguistic models are explanatory about semantic processing in the brain's language network [60–64]. Combining the grounded language model with human behavioral and neural data may elucidate how the language network interacts with distributed sensorimotor areas for semantic processing [65].

It is natural to extend this study by incorporating other sensory input [66, 67] and further ground language learning in action [68–71] and emotion [72]. This study also leaves an open question as to whether grounding should occur at an early or late stage of natural language processing, which awaits further exploration and evaluation. For comprehensive modeling of language grounding, it is desirable to expose an agent to a naturalistic and multi-sensory environment and to engage interactive actions that allows the agent to learn knowledge in the physical world, like how humans learn language.

References

[1] J Ridley Stroop. Studies of interference in serial verbal reactions. *Journal of experimental psychology*, 18 (6):643, 1935.

[2] Friedemann Pulvermüller. How neurons make meaning: brain mechanisms for embodied and abstract-symbolic semantics. *Trends in cognitive sciences*, 17(9):458–470, 2013.

[3] Fritz Günther, Tri Nguyen, Lu Chen, Carolin Dudschig, Barbara Kaup, and Arthur M Glenberg. Immediate sensorimotor grounding of novel concepts learned from language alone. *Journal of Memory and Language*, 115:104172, 2020.

[4] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26: 3111–3119, 2013.

[5] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*, 27:3104–3112, 2014.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[7] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. 33:1877–1901, 2020.

[8] Guy Emerson. What are the goals of distributional semantics? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7436–7453, 2020.

[9] John R Searle et al. Minds, brains, and programs. *The Turing Test: Verbal Behaviour as the Hallmark of Intelligence*, pages 201–224, 1980.

[10] Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346, 1990.

[11] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[12] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6700–6709, 2019.

[13] Alex Martin. Grapes—grounding representations in action, perception, and emotion systems: How object properties and categories are represented in the human brain. *Psychonomic bulletin & review*, 23(4): 979–990, 2016.

[14] Lawrence W Barsalou. Grounded cognition. *Annu. Rev. Psychol.*, 59:617–645, 2008.

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

[16] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[17] Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, et al. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, 2020.

[18] Arthur M Glenberg and David A Robertson. Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of memory and language*, 43(3):379–401, 2000.

[19] Carina Silberer and Mirella Lapata. Grounded models of semantic representation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1423–1433, 2012.

[20] Stephen Roller and Sabine Schulte Im Walde. A multimodal lda model integrating textual, cognitive and visual modalities. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1146–1157, 2013.

[21] Elia Bruni, Nam-Khanh Tran, and Marco Baroni. Multimodal distributional semantics. *Journal of artificial intelligence research*, 49:1–47, 2014.

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

[23] Eloi Zablocki, Benjamin Piwowarski, Laure Soulier, and Patrick Gallinari. Learning multi-modal word representation grounded in visual context. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[24] Angeliki Lazaridou, Marco Baroni, et al. Combining language and vision with a multimodal skip-gram model. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 153–163, 2015.

[25] Douwe Kiela, Alexis Conneau, Allan Jabri, and Maximilian Nickel. Learning visually grounded sentence representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 408–418, 2018.

[26] Xisen Jin, Junyi Du, Arka Sadhu, Ram Nevatia, and Xiang Ren. Visually grounded continual learning of compositional phrases. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2018–2029, 2020.

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

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30:5998–6008, 2017.

[29] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5103–5114, 2019.

[30] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265*, 2019.

[31] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*, 2019.

[32] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. 2019.

[33] Suzanna Becker and Geoffrey E Hinton. Self-organizing neural network that discovers surfaces in random-dot stereograms. *Nature*, 355(6356):161–163, 1992.

[34] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a" siamese" time delay neural network. *Advances in neural information processing systems*, 6: 737–744, 1993.

[35] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

[36] David Harwath, Adria Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James Glass. Jointly discovering visual objects and spoken words from raw sensory input. In *Proceedings of the European conference on computer vision (ECCV)*, pages 649–665, 2018.

[37] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv preprint arXiv:2102.05918*, 2021.

[38] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2019.

[39] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

[40] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European conference on computer vision*, pages 852–869. Springer, 2016.

[41] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. *Advances in Neural Information Processing Systems*, 30:4967–4976, 2017.

[42] Xinlei Chen, Li-Jia Li, Li Fei-Fei, and Abhinav Gupta. Iterative visual reasoning beyond convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7239–7248, 2018.

[43] Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. Relation-aware graph attention network for visual question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10313–10322, 2019.

[44] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4654–4662, 2019.

[45] Huda Hakami, Danushka Bollegala, and Hayashi Kohei. Why pairdiff works?–a mathematical analysis of bilinear relational compositional operators for analogy detection. *arXiv preprint arXiv:1709.06673*, 2017.

[46] Tomáš Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pages 746–751, 2013.

[47] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. In *Neural Information Processing Systems (NIPS)*, pages 1–9, 2013.

[48] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. Knowledge graph embedding by translating on hyperplanes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 28, 2014.

[49] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. *arXiv preprint arXiv:1412.6575*, 2014.

[50] Maximilian Nickel, Lorenzo Rosasco, and Tomaso Poggio. Holographic embeddings of knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

[51] Gabriel Ilharco, Rowan Zellers, Ali Farhadi, and Hannaneh Hajishirzi. Probing text models for common ground with visual representations. *arXiv preprint arXiv:2005.00619*, 2020.

[52] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[53] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pages 213–229. Springer, 2020.

[54] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[55] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[56] Lütfi Kerem Şenel, Ihsan Utlu, Veysel Yücesoy, Aykut Koc, and Tolga Cukur. Semantic structure and interpretability of word embeddings. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(10):1769–1779, 2018.

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

[58] Lucy Li and Jon Gauthier. Are distributional representations ready for the real world? evaluating word vectors for grounded perceptual meaning. In *Proceedings of the First Workshop on Language Grounding for Robotics*, pages 76–85, 2017.

[59] Barry J Devereux, Lorraine K Tyler, Jeroen Geertzen, and Billi Randall. The centre for speech, language and the brain (cslb) concept property norms. *Behavior research methods*, 46(4):1119–1127, 2014.

[60] Tom M Mitchell, Svetlana V Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L Malave, Robert A Mason, and Marcel Adam Just. Predicting human brain activity associated with the meanings of nouns. *science*, 320(5880):1191–1195, 2008.

[61] Alexander G Huth, Wendy A De Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458, 2016.

[62] Luana Bulat, Stephen Clark, and Ekaterina Shutova. 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*, pages 1081–1091, 2017.

[63] Francisco Pereira, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel J Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. Toward a universal decoder of linguistic meaning from brain activation. *Nature communications*, 9(1):1–13, 2018.

[64] Yizhen Zhang, Kuan Han, Robert Worth, and Zhongming Liu. Connecting concepts in the brain by mapping cortical representations of semantic relations. *Nature communications*, 11(1):1–13, 2020.

[65] Stefan Heinrich, Yuan Yao, Tobias Hinz, Zhiyuan Liu, Thomas Hummel, Matthias Kerzel, Cornelius Weber, and Stefan Wermter. Crossmodal language grounding in an embodied neurocognitive model. *Frontiers in Neurorobotics*, 14, 2020.

[66] Douwe Kiela, Luana Bulat, and Stephen Clark. 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)*, pages 231–236, 2015.

[67] Douwe Kiela and Stephen Clark. Multi-and cross-modal semantics beyond vision: Grounding in auditory perception. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2461–2470, 2015.

[68] Haonan Yu, Haichao Zhang, and Wei Xu. Interactive grounded language acquisition and generalization in a 2d world. *arXiv preprint arXiv:1802.01433*, 2018.

[69] Joyce Y Chai, Qiaozi Gao, Lanbo She, Shaohua Yang, Sari Saba-Sadiya, and Guangyue Xu. Language to action: Towards interactive task learning with physical agents. In *IJCAI*, pages 2–9, 2018.

[70] Rohan Paul, Jacob Arkin, Nicholas Roy, and Thomas M Howard. Efficient grounding of abstract spatial concepts for natural language interaction with robot manipulators. 2016.

[71] Corey Lynch and Pierre Sermanet. Grounding language in play. *arXiv preprint arXiv:2005.07648*, 2020.

[72] Armand S Rotaru and Gabriella Vigliocco. Constructing semantic models from words, images, and emojis. *Cognitive science*, 44(4):e12830, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 5.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We include the code and instructions needed to reproduce the main experimental results in the supplementary material. The dataset used in this study is publicly available.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We include the most important training details in Section 3.3. Other details are elaborated in the supplementary material.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We include the total amount of computational resources in the supplementary material.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] See Section 3.3 and Section 4
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]