Probing Contextual Language Models for Common Ground with Visual Representations

Gabriel Ilharco **Rowan Zellers** **Ali Farhadi** **Hannaneh Hajishirzi**
Paul G. Allen School of Computer Science & Engineering
University of Washington
{gamaga,rowanz,ali,hannaneh}@cs.washington.edu

Abstract

The success of large-scale contextual language models has attracted great interest in probing what is encoded in their representations. In this work, we consider a new question: to what extent contextual representations of concrete nouns are aligned with corresponding visual representations? We design a probing model that evaluates how effective are text-only representations in distinguishing between matching and non-matching visual representations. Our findings show that language representations alone provide a strong signal for retrieving image patches from the correct object categories. Moreover, they are effective in retrieving specific instances of image patches; textual context plays an important role in this process. Visually grounded language models slightly outperform text-only language models in instance retrieval, but greatly under-perform humans. We hope our analyses inspire future research in understanding and improving the visual capabilities of language models.

1 Introduction

Contextual language models trained on text-only corpora are prevalent in recent natural language processing (NLP) literature (Devlin et al., 2019; Liu et al., 2019b; Lan et al., 2019; Raffel et al., 2019). Understanding what their representations encode has been the goal of a number of recent studies (Belinkov and Glass, 2019; Rogers et al., 2020). Yet, much is left to be understood about whether—or to what extent—these models can encode visual information.

We study this problem in the context of language grounding (Searle et al., 1984; Harnad, 1990; McClelland et al., 2019; Bisk et al., 2020; Bender and Koller, 2020), empirically investigating whether text-only representations can naturally be connected to the visual domain, without explicit visual supervision in pre-training.

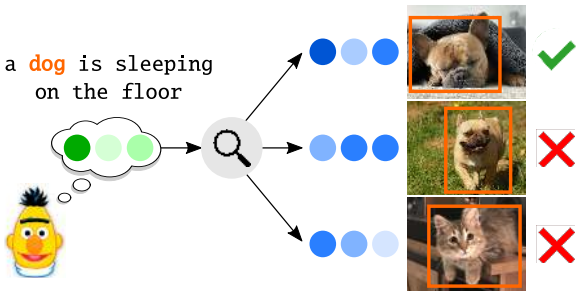

Figure 1: We introduce a probing mechanism that learns a mapping from contextual language representations to visual features. For a number of contextual language models, we evaluate how useful their representations are for retrieving matching image patches.

We argue that *context* plays a significant role in this investigation. In language, the ability to form context-dependent representations has shown to be crucial in designing pre-trained language models (Peters et al., 2018; Devlin et al., 2019). This is even more important for studying grounding since many visual properties depend strongly on context (Sadeghi and Farhadi, 2011). For instance, a "`flying bat`" shares very few visual similarities with a "`baseball bat`"; likewise, a "`dog sleeping`" looks different from a "`dog running`". While alignments between language representations and visual attributes have attracted past interest (Leong and Mihalcea, 2011; Lazaridou et al., 2014, 2015; Lucy and Gauthier, 2017; Collell Talleda et al., 2017), the role of context has been previously overlooked, leaving many open questions about what visual information contextual language representations encode.

In this work, we introduce a method for empirically *probing* contextual language representations and their relation to the visual domain. In general, probing examines properties for which the models are not designed to predict, but can be encoded in their representations (Shi et al., 2016; Rogers et al.,

Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 5367–5377
June 6–11, 2021.

Figure 2: Examples of retrieved image patches from text-only representations using our probe. All shown images are retrieved from MS-COCO (Lin et al., 2014), using representations from BERT base. Importantly, these object categories (e.g. kite) are previously *unseen* by our probe. On the bottom rows, we show examples of the influence of context in retrieval: while all retrieved image patches belong to the correct object category, cat, more descriptive contexts allow more accurate retrieval at the instance level.

2020). Here, our probe is a lightweight model trained to map language representations of concrete objects to corresponding visual representations. The probe (illustrated in Figure 1) measures whether language representations can be used to give higher scores to matching visual representations compared to mismatched ones.

Textual and visual representations are collected from image captioning data, where we find pairs of concrete words (e.g. cat or kite) and their corresponding image patches. The probe is trained using a contrastive loss (Oord et al., 2018) that gauges the mutual information between the language and visual representations. Given text-only representations of an unseen object category, the trained probe is evaluated by retrieving corresponding image patches for categories it has never seen during training. Qualitative examples can be found in Figure 2.

We examine representations from a number of contextual language models including BERT, RoBERTa, ALBERT and T5 (Devlin et al., 2019; Liu et al., 2019b; Lan et al., 2019; Raffel et al., 2019). For all of them, we find that interesting mappings can be learned from language to visual representations, as illustrated in Figure 2. In particular, using its top-5 predictions, BERT representations retrieve the correctly paired visual instance 36% of the time, strongly outperforming non-contextual language models (e.g., GloVe (Pennington et al., 2014)). Moreover, for all examined models, image patches of the correct object category are retrieved with a recall of 84-90%. Our experiments are backed by a control task where visual representations are intentionally mismatched with their textual counterparts. Retrieval performance drops substantially in these settings, attesting the selectivity of our probe.

Moreover, we measure the impact of context on retrieval at the instance level. Contextual models substantially outperform non-contextual embeddings, but this difference disappears as context is gradually hidden from contextual models. When the context includes adjectives directly associated with the noun being inspected, we find significantly better instance retrieval performance.

Finally, we investigate a number of grounded language models—such as LXMERT and VILBERT (Tan and Bansal, 2019; Lu et al., 2019, 2020)—that see visual data in training, finding them to slightly outperform text-only models. Contrasting the learned mappings with human judgment, the examined visually grounded language models significantly underperform human subjects, exposing much room for future improvement.

2 Related Work

What is encoded in language representations? Understanding what information NLP models encode has attracted great interest in recent years

(Rogers et al., 2020). From factual (Petroni et al., 2019; Jawahar et al., 2019; Roberts et al., 2020) to linguistic (Conneau et al., 2018; Liu et al., 2019a; Talmor et al., 2019) and commonsense (Forbes et al., 2019) knowledge, a wide set of properties have been previously analysed. We refer to Belinkov and Glass (2019) and Rogers et al. (2020) for a more comprehensive literature review. A common approach, often used for inspecting contextual models, is probing (Shi et al., 2016; Adi et al., 2016; Conneau et al., 2018; Hewitt and Liang, 2019). In short, it consists of using supervised models to predict properties not directly inferred by the models. Probing is typically used in settings were discrete, linguistic annotations such as parts of speech are available. Our approach differs from previous work in both scope and methodology, using a probe to measure similarities with continuous, visual representations. Closer to our goal of better understanding grounding is the work of Cao et al. (2020), that design probes for examining multi-modal models. In contrast, our work examines text-only models and does not rely on their ability to process images.

Language grounding. A widely investigated research direction aims to connect natural language to the physical world (Bisk et al., 2020; McClelland et al., 2019; Tan and Bansal, 2019; Lu et al., 2019, 2020; Chen et al., 2020; Li et al., 2020; Tan and Bansal, 2020). This is typically done through training and evaluating models in tasks and datasets where both images and text are used, such as visual question answering (Antol et al., 2015; Hudson and Manning, 2019). A number of previous work have investigated mappings between language and visual representations or mappings from both to a shared space. Leong and Mihalcea (2011) investigate semantic similarities between words and images through a joint latent space, finding a positive correlation with human rated similarities. Similarly, Silberer and Lapata (2014) builds multi-modal representations by using stacked autoencoders. Socher et al. (2013) and Lazaridou et al. (2014) show that a shared latent space allows for zero-shot learning, demonstrating some generalization to previously unseen objects. Lazaridou et al. (2015) construct grounded word representations by exposing them to aligned visual features at training time. Lucy and Gauthier (2017) investigate how well word representations can predict perceptual and conceptual features, showing that a number of such features are not adequately predicted. Collell Talleda et al. (2017) uses word embeddings to create a mapping from language to visual features, using its outputs to build multimodal representations. While our conclusions are generally aligned, our work differs from these in two important ways. Firstly, previous work studies context-independent word representations, while our method allows analysing language representations that depend on the context they are used in. We use this to examine a number of trained contextual language models. Secondly, while most previous work uses these mappings for building better grounded representations—often training the language models in the process—our work focuses on using them as a tool for inspecting already trained models, without modifying them.

Zero-shot detection. Recent work attempts to build object detectors that generalize to unseen object categories, by conditioning the predictions on word embeddings of the class (Rahman et al., 2018; Demirel et al., 2018), visual attributes (Demirel et al., 2018; Zhu et al., 2019; Mao et al., 2020) or text descriptions (Li et al., 2019). In our work, we use language representations of words in context (captions) as inputs. More fundamentally, although our experiments on unseen object categories can be used for zero-shot detection, we differ from previous work in motivation, which translates to further experimental differences. Given our goal to analyse already trained models (as opposed to learning a generalizable object detector), we train nothing apart from a lightweight probe in our analyses.

3 Probing contextual representations

Our main goal is to characterize the relation between contextual language representations and the visual domain. We first describe how language and visual representations of concrete concepts can be collected from image captioning datasets (§3.1). Next, we design a probe that examines the relation between these representations, learning a mapping from language to visual representations (§3.2). An overview is illustrated in Figure 3.

3.1 Collecting data

At the center of our analysis are contextual representations of visually observable nouns, which we refer to as *object categories*. Here, we describe how pairs of matching language and visual representations (ℓ, v) are collected from image captioning datasets.

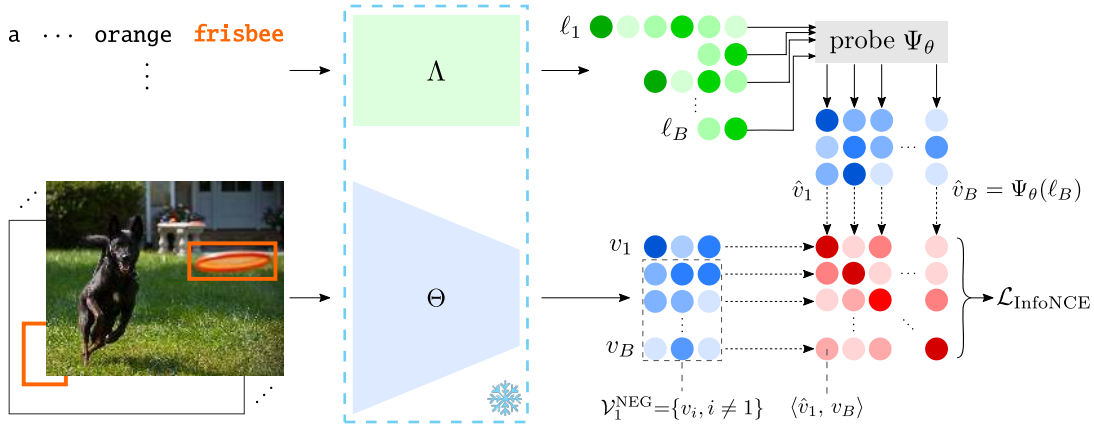

Figure 3: An overview of the proposed probing procedure. Frozen language and vision models (Λ and Θ) extract representations from matching pairs of words in text and objects in images. A probe Ψ_θ is trained to map representations from text (green) to visual (blue) domains while maximally preserving mutual information. For a given language representation ℓ_i, the loss (Equation 1) drives the probe's outputs $\hat{v}_i{=}\Psi_\theta(\ell_i)$ to be maximally useful for finding the aligned visual representation v_i given all other visual representations in the batch ($\mathcal{V}_i^{\text{NEG}} = v_j, i \neq j$). For such, only the pair-wise dot products $\langle \hat{v}_i, v_j \rangle$ are required (red).

Language representations (ℓ) are extracted from image captions. To accommodate recent language models and tokenizers, we allow such representations to be contextual and have variable length,[1] where each element in ℓ has a fixed dimension d_L. The length of the representations ℓ for each object category is determined by the tokenizer. We treat a model that extracts representations from text as a function Λ that maps a string o (here, object categories) in a larger textual context c (here, captions) to the representation $\ell = \Lambda(o \mid c)$. This formalism also encompasses non-contextual embeddings, with $\Lambda(o \mid c) = \Lambda(o)$.

Visual representations (v) are extracted from objects in images using a trained object detection model Θ. For simplicity, we use $v = \Theta(o \mid i)$ to refer to the extracted features corresponding to the detected object from image i that is both 1) classified as a member of object category o and 2) assigned the highest confidence by the model among those. Visual representations $\Theta(o \mid i)$ have fixed dimensions d_V.

Paired data (ℓ, v) with aligned representations is collected from an image captioning dataset with paired captions c and images i. For each image i, and each object o detected by the object detector Θ, if o appears in some associated caption c, we include the pair $(\ell = \Lambda(o \mid c), v = \Theta(o \mid i))$. To avoid having multiple pairs (ℓ, v) associated with the same visual instance, we ensure that at most one pair (ℓ, v) per object category in each image is included. In this work, we use the 1600 object categories from Faster R-CNN (Ren et al., 2015) trained on Visual Genome (Krishna et al., 2017).

[1] Conforming with sub-word tokenizers or multi-word expressions such as `fire extinguisher`.

3.2 Probing representations

At a high level, language representations are inspected via a shallow neural probing model (Figure 3). In training, the probe learns a mapping from language to visual representations (§3.2.1). We then evaluate the quality of these mappings by measuring how well they can be used to retrieve matching image patches (§3.2.2).

3.2.1 Training the probe

The probe is optimized to maximally preserve the mutual information between the distributions of language and visual representations. This is done via InfoNCE (Oord et al., 2018) (Equation 1), a loss function commonly used for retrieval and contrastive learning (Le-Khac et al., 2020). We note the mutual information is a bottleneck on how well two random variables can be mapped to one another, given its relation to conditional entropy. In training, the probe Ψ_θ with parameters θ takes inputs ℓ and estimates visual representations $\hat{v} = \Psi_\theta(\ell)$ with the same dimensionality d_V as the corresponding visual representations v. For each pair (ℓ, v), this loss relies on a set of distractors $\mathcal{V}_\ell^{\text{NEG}}$, containing visual representations which are *not* aligned with the language representations ℓ. The representations in $\mathcal{V}_\ell^{\text{NEG}}$ are used

for contrastive learning and are drawn from the same visual model, using different objects or images. Minimizing this loss drives the dot product $\langle \Psi_\theta(\ell), u \rangle$ to be maximal for $u = v$ and small for all $u \in \mathcal{V}_\ell^{\text{NEG}}$. In other words, training pushes the estimates $\hat{v} = \Psi_\theta(\ell)$ to be maximally useful in discerning between positive and negative visual pairings.

$$\mathcal{L}_{\text{InfoNCE}} = -\mathbb{E}_\ell \left[\log \frac{e^{\langle \Psi_\theta(\ell), v_\ell \rangle}}{\sum_{v' \in \{v\} \bigcup \mathcal{V}_\ell^{\text{NEG}}} e^{\langle \Psi_\theta(\ell), v' \rangle}} \right] \quad (1)$$

In practice, the expectation in Equation 1 is estimated over a batch of size B with samples of aligned language and visual representations $((\ell_1, v_1), \ldots, (\ell_B, v_B))$. For efficiency, we use other visual representations in the batch as distractors for a given representation ($\mathcal{V}_i^{\text{NEG}} = \{v_j, j \neq i\}$). Thus, only the dot products $\langle \hat{v}_i = \Psi_\theta(\ell_i), v_j \rangle$ are needed to calculate the loss, as illustrated in Figure 3. Importantly, we note that the models used to extract representations are not trained or changed in any way during the probing procedure.

3.2.2 Evaluation procedure

For evaluation, we compute recall in retrieving image patches given objects in text, using new pairs of language and visual representations from unseen images and captions. Consider the set of all collected visual representations for evaluation, $\mathcal{V}$. For each language representation ℓ, we use the trained probe to generate our estimate $\hat{v} = \Psi_\theta(\ell)$, and find the instances $v' \in \mathcal{V}$ that maximize the dot product $\langle \hat{v}, v' \rangle$. Given an integer k, we consider recall at k at both instance and category levels. Formally:

Instance Recall (IR@k) measures how frequently the correct visual instance is retrieved. More precisely, it is the fraction of pairs (ℓ, v) where the instance v is in the top-k visual representations retrieved from $\hat{v} = \Psi_\theta(\ell)$.

Category Recall (CR@k) measures how frequently instances of the correct object category are retrieved. More precisely, it is the fraction of pairs $(\ell, v = \Theta(o \mid i))$ where any of the top-k retrieved visual representations $v' = \Theta(o' \mid i')$ belongs to the same object category as v (i.e. $o' = o$).

Higher IR and CR scores indicate better performance and, by definition, CR@k cannot be smaller than IR@k. These metrics form the basis of our evaluation, and we take multiple steps to promote experimental integrity. Learned mappings are evaluated in two scenarios, where pairs (ℓ, v) are collected using object categories either *seen* or *unseen* by the probe during training. The later is the focus of the majority of our experiments. For both scenarios, images and captions have no intersection with those used in training. Further, we create multiple *seen/unseen* splits from our data, training and testing on each split. We then report average and standard deviation of the recall scores across 5 splits.

4 Experimental settings

4.1 Language models

The majority of examined models are contextual representation models based on the transformer architecture (Vaswani et al., 2017) trained on text-only data. We examine the *base* ($d_L = 768$) and *large* ($d_L = 1024$) versions of BERT uncased, RoBERTa, ALBERT and T5 (Devlin et al., 2019; Liu et al., 2019b; Lan et al., 2019; Raffel et al., 2019). For T5, we also examine the *small* version, with $d_L = 512$. For all these models, we use pretrained weights from the HuggingFace Transformers library (Wolf et al., 2020)[2], and use representations from the last layer. Additionally, we inspect non-contextual representations using GloVe embeddings (Pennington et al., 2014), using embeddings trained on 840 billion tokens of web data, with $d_L = 300$ and a vocabulary size of 2.2 million.[3]

4.2 Vision models

As is common practice in natural language grounding literature (Anderson et al., 2018; Tan and Bansal, 2019; Su et al., 2020; Lu et al., 2020), we use a Faster R-CNN model (Ren et al., 2015) trained on Visual Genome (Krishna et al., 2017) to extract visual features with $d_V = 2048$. We use the trained network provided by Anderson et al. (2018)[4], and do not fine-tune during probe training.

4.3 Data

We collect representations from two image captioning datasets, Flickr30k (Young et al., 2014), with over 150 thousand captions and 30 thousand images, and MS-COCO (Lin et al., 2014), with 600 thousand captions and 120 thousand images

[2] https://github.com/huggingface/transformers
[3] https://nlp.stanford.edu/projects/glove/
[4] https://github.com/peteanderson80/bottom-up-attention

in English. The larger MS-COCO is the focus of the majority of our experiments. We build *disjoint* training, validation and test sets from the aggregated training and validation image captions. To examine generalization to new objects, we test on representations from both *seen* or *unseen* object categories, built from images and captions not present in the training data. From the 1600 object categories of our object detector, we use 1400 chosen at random for training and *seen* evaluation. The remaining 200 are reserved for *unseen* evaluation. Furthermore, we train and test our probe 5 times, each with a different 1400/200 split of the object categories. For each object category split, we build validation and test sets with sizes proportional to the number of object categories present: *seen* test sets contain 7000 representation pairs and *unseen* test sets contain 1000 pairs. The validation sets used for development consists of *seen* object categories, with the same size as the *seen* test sets. All remaining data is used for training.

4.4 Control task

Contrasting the probe performance with a control task is central to probing (Hewitt and Liang, 2019). We follow this practice by learning in a control task where representations are mapped to *permuted* visual representations. More precisely, we replace each visual representation $v = \Theta(o \mid i)$ with another $v' = \Theta(o' \mid i')$ chosen at random from an object category $o' = f(o)$ that depends on the original object category o. Here, f dictates a random permutation of the object categories. For instance, visual representations of the original category `cat` are replaced with representations from a second category `dog`; representations from the category `dog` are replaced by those from `tree`, and so on.

4.5 Implementation and hyper-parameters

Our probe consists of a shallow neural model. To process the naturally sequential language representations ℓ, we use a single-layered model with LSTM cells (Hochreiter and Schmidhuber, 1997) with 256 hidden units and only unidirectional connections. The outputs are then projected by a linear layer to the visual space. The probe is trained using Adam optimizer (Kingma and Ba, 2015) with a learning rate of 0.0005, weight decay of 0.0005 and default remaining coefficients (β_1=0.9 β_2=0.999 and ϵ=10^{-8}). We train with a batch size of 3072, for a total of 5 epochs on one GPU.

#	Experiment	IR@1	IR@5	CR@1
0	Random	0.1 ± 0.1	0.5 ± 0.2	6.0 ± 2.0
1	Control	0.0 ± 0.0	0.4 ± 0.2	3.0 ± 1.4
2	GloVe	5.1 ± 0.5	18.5 ± 1.4	87.3 ± 3.5
3	BERT base	12.0 ± 1.0	36.0 ± 0.9	88.1 ± 2.4
4	BERT large	11.6 ± 0.7	34.9 ± 2.6	89.3 ± 2.4
5	RoBERTa base	11.6 ± 0.3	34.4 ± 2.2	90.4 ± 0.6
6	RoBERTa large	10.9 ± 1.1	32.8 ± 2.5	88.7 ± 3.2
7	ALBERT base	8.7 ± 0.2	28.8 ± 1.6	84.4 ± 2.1
8	ALBERT large	9.4 ± 1.0	28.8 ± 2.3	84.2 ± 4.2
9	T5 small	10.1 ± 0.7	32.9 ± 1.5	87.2 ± 4.1
10	T5 base	10.8 ± 0.8	33.3 ± 2.3	85.3 ± 2.8
11	T5 large	11.8 ± 0.5	34.7 ± 2.1	87.2 ± 2.4

Table 1: Average instance recall (IR@k) and category recall (CR@k) for test sets with *unseen* object categories. For each model, we train and evaluate 5 times, using different sets of object categories seen in training. Unlike the control task with permuted representations, mappings learned from sensible representations generalize well to unseen object categories.

5 Results and discussion

At a high level, our experiments show that i) language representations are strong signals for choosing between different visual features both at the instance and category levels; ii) context is largely helpful for instance retrieval; iii) InfoNCE works better than other studied losses, and some consistency is found across datasets; iv) visually grounded models outperform text-only models; v) all models lag greatly behind human performance. We provide further details in §5.1-5.3.

5.1 Retrieval results

Table 1 summarizes instance and category retrieval performance for different language models and control experiments, using test data with *unseen* object categories. Our results indicate that language representations alone are strong signals for predicting visual features: for all examined language models, recall scores are significantly better than random and control. Qualitative results can be found in Figure 2. We note that category recall scores are significantly higher than instance recall. This is reasonable since there are many more positive alignments at the category level. Compared to other inspected models, BERT base shows the best results for instance retrieval, and will be the focus of further analyses.

Contrasting the performance of non-contextual representations from GloVe with that of contextual models shows that context considerably affects

#	Experiment	IR@1	IR@5	CR@1
0	Random	0.1 ± 0.1	0.1 ± 0.1	1.2 ± 0.1
1	Control	1.6 ± 0.1	7.8 ± 0.6	41.3 ± 5.6
2	BERT base	14.9 ± 0.3	43.4 ± 0.8	90.4 ± 0.4

Table 2: Average instance recall (IR@k) and category recall (CR@k) for test sets with *seen* object categories.

Loss function	IR@1	IR@5	CR@1
MSE	3.0 ± 0.3	12.1 ± 1.3	57.5 ± 8.7
Neg. cosine sim.	6.9 ± 0.7	23.4 ± 1.3	75.1 ± 6.3
Triplet loss	8.4 ± 0.6	28.8 ± 0.9	81.7 ± 3.6
InfoNCE	12.0 ± 1.0	36.0 ± 0.9	88.1 ± 2.4

Table 3: Comparison in retrieval performance on *unseen* object categories for different training losses, using representations from BERT base. InfoNCE yields better results than other loss functions.

instance recall. For instance, GloVe and BERT base yield 5.1% to 12.0% IR@1, respectively. This gap is sensible, since a non-contextual representation should not be able to discern between distinct image patches depicting the same object category. While still lagging behind a number of contextual representations, we observe strong category recall for GloVe, which we hypothesize is due to the ease in predicting the correct output category since input representations are fixed, independently of context. We further explore the role of context in §5.3.

Moreover, Table 2 shows performance on test sets with *seen* object categories. Comparing with Table 1, BERT representations show good generalization to unseen object categories. This generalization is consistent with previous observations on zero-shot experiments, using non-contextual word embeddings (Lazaridou et al., 2014).

Finally, our results attest to the selectivity of the probe: for the control task with permuted representations (Tables 1 and 2, Row 1), substantially lower performance is found. This gap is particularly high for *unseen* object categories, where only sensibly paired representations perform better than chance.

5.2 Ablations

Loss ablations. In addition to InfoNCE, we ablate on 3 other loss functions: mean squared error (MSE), negative cosine similarity, and triplet loss[5]. The results for unseen object categories are summarized in Table 3: while all losses yield better than random results, InfoNCE performs the best. This validates the theoretical intuition that InfoNCE would be advantageous, as it allows for directly optimizing the probe to maximally preserve the mutual information between the representations, a bottleneck on the remaining entropy after the mapping.

[5] $\mathcal{L}_{trip} = \mathbb{E}_\ell[\max(\delta_{\ell,v'_\ell} - \delta_{\ell,v_\ell} + \alpha, 0)]$, where the margin α is set to 1.0, $v' \in \mathcal{V}^{\text{NEG}}$ and $\delta_{\ell,v} = \cos(\Psi_\theta(\ell), v_\ell)$.

Dataset	# Images / # Captions	IR@1	CR@1
MS-COCO	120k / 600k	12.0 ± 1.0	88.1 ± 2.4
Flickr30k	30k / 150k	9.8 ± 0.9	85.6 ± 3.4

Table 4: Comparison for different datasets in retrieval performance of *unseen* object categories with representations from BERT base. Despite large differences in size, results indicate consistency across datasets.

Data ablations. In addition to MS-COCO, which is the used for the majority of our experiments, we show results with data collected from the smaller Flickr30k. We report the test retrieval performance for unseen object categories using representations from BERT base in Table 4. These results indicate consistency across the datasets, despite their considerable difference in size.

5.3 Analyses

Influence of context. We study whether the gap in instance retrieval performance from GloVe and BERT comes from the use of context or intrinsic differences of these models. This is explored by measuring how instance recall varies as we probabilistically mask out context tokens in the captions at different rates. As shown in Figure 4, performance drops substantially as more tokens are masked; in the limit where only the object tokens remain (i.e. the fraction of context masked is 1.0), BERT's representations perform marginally worse than the non-contextual GloVe embeddings.

Figure 5 compares instance-level retrieval accuracy for representations when objects have none or at least one adjective associated with them, as processed by the dependency parser from AllenNLP library (Gardner et al., 2018). These adjectives commonly include colors (e.g. `white`, `black`) and sizes (e.g. `big`, `small`), indicating contextual information. The results show clear gains in instance recall when objects are accompanied by adjectives, confirming that context enables more accurate retrieval. We refer back to Figure 2 for qualitative results on the influence of context.

Grounded language models. We further inspect representations from several grounded language

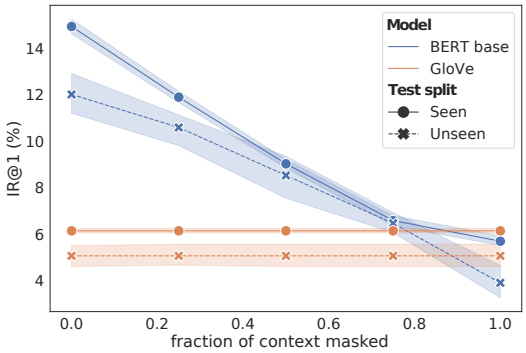

Figure 4: Instance recall as context tokens are progressively masked out. Retrieval performance for BERT quickly degrades as higher proportions of the context are masked.

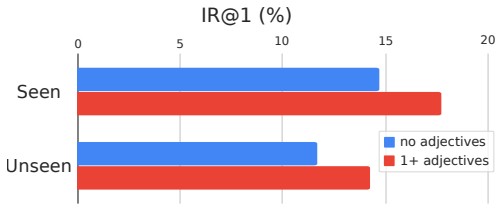

Figure 5: More descriptive contexts enable more accurate retrieval. In the plot, we show instance recall at 1 when object categories are or are not accompanied by adjectives, using representations from BERT base.

models, namely LXMERT, VL-BERT (base and large) and VILBERT-MT (Tan and Bansal, 2019; Su et al., 2020; Lu et al., 2019, 2020)). While these models typically process visual and textual inputs jointly, we adapt them to include only the language branches, restricting attention to the text inputs. For all these models, we use the code and weights made public by the authors.[6] The results, summarized in Table 5, show that grounded models slightly outperform the ungrounded BERT base. At the category level, we see small relative differences in performance between grounded and ungrounded models. At the instance level, the relative improvement is higher, especially for VILBERT-MT, while still much lower than human performance as shown in the next experiment.

Human performance. Finally, we contrast the examined models with human performance in retrieving visual patches given words in sentences. Such a comparison helps disentangling the quality of the learned mappings with possible incidental matches, i.e., language representations with more than one positive visual match. As they are also affected by these artifacts, human subjects offer a sensible point of comparison. In virtue of the limited human attention, we evaluate on a reduced test set with unseen object categories, randomly sampling 100 data points from it. For each object in a sentence, subjects are presented with 100 image patches and asked to choose the closest match. We collect over 1000 annotations from 17 in-house annotators, with at least 30 annotations each. Our results are shown in Table 6. On the same test set, we find a large gap from learned mappings for both grounded and ungrounded models to human performance, exposing much room for improvement.

[6] github.com/airsplay/lxmert; github.com/jackroos/VL-BERT; github.com/facebookresearch/vilbert-multi-task

Model	IR@1	IR@5	CR@1
BERT base	12.0 ± 1.0	36.0 ± 0.9	88.1 ± 2.4
LXMERT	13.7 ± 1.0	39.2 ± 2.5	90.3 ± 1.2
VL-BERT base	12.5 ± 1.0	37.6 ± 1.1	88.7 ± 1.4
VL-BERT large	12.6 ± 1.1	37.5 ± 2.4	88.7 ± 2.3
VILBERT-MT	15.4 ± 1.2	42.4 ± 2.7	90.8 ± 1.9

Table 5: Retrieval performance for *unseen* object categories, using representations from BERT and a number of grounded language models.

Chance	BERT base	VILBERT-MT	Human
1%	43%	53%	76%

Table 6: A sizable gap in instance recall (IR@1) is seen by comparing the performance of humans and the examined models in a reduced test set with 100 samples.

6 Conclusion

Understanding the similarities between language and visual representations has important implications on the models, training paradigms and benchmarks we design. We introduced a method for empirically measuring the relation between contextual language representations and corresponding visual features. We found contextual language models to be useful—while far from human subjects—in discerning between different visual representations. Moreover, we explored how these results are influenced by context, loss functions, datasets and explicit grounding during training. Altogether, we hope that our new methodological and practical insights foster further research in both understanding the natural connections between language and visual representations and designing more effective models at the intersection the two modalities.

Acknowledgements

This research was supported by the grants from ONR N00014-18-1-2826, DARPA N66001-19-2-4031, 67102239, NSF III-1703166, IIS-1652052, IIS-17303166, and an Allen Distinguished Investigator Award and a Sloan Fellowship. Authors would also like to thank Raymond J. Mooney and members of the UW-NLP, H2Lab and RAIVN Lab at the University of Washington for their valuable feedback and comments.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2016. Fine-grained analysis of sentence embeddings using auxiliary prediction tasks. *arXiv preprint arXiv:1608.04207*.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*.

Emily M. Bender and Alexander Koller. 2020. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. 2020. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. In *Proceedings of the 2020 European Conference on Computer Vision*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Learning universal image-text representations. In *Proceedings of the 2020 European Conference on Computer Vision*.

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

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Berkan Demirel, Ramazan Gokberk Cinbis, and Nazli Ikizler-Cinbis. 2018. Zero-shot object detection by hybrid region embedding. *arXiv preprint arXiv:1805.06157*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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*. Association for Computational Linguistics.

Maxwell Forbes, Ari Holtzman, and Yejin Choi. 2019. Do neural language representations learn physical commonsense? *Proceedings of the 41st Annual Conference of the Cognitive Science Society*.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*. Association for Computational Linguistics.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*.

J. Hewitt and P. Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of Empirical Methods in Natural Language Processing*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*.

Drew A Hudson and Christopher D Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *Proceedings of the 2015 International Conference for Learning Representations*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. In *Proceedings of the 2020 International Conference on Learning Representations*.

Angeliki Lazaridou, Elia Bruni, and Marco Baroni. 2014. Is this a wampimuk? cross-modal mapping between distributional semantics and the visual world. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Angeliki Lazaridou, Nghia The Pham, and Marco Baroni. 2015. Combining language and vision with a multimodal skip-gram model. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

P. H. Le-Khac, G. Healy, and A. F. Smeaton. 2020. Contrastive representation learning: A framework and review. *IEEE Access*.

Chee Wee Leong and Rada Mihalcea. 2011. Measuring the semantic relatedness between words and images. In *Proceedings of the Ninth International Conference on Computational Semantics (IWCS 2011)*.

Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Proceedings of the 2020 European Conference on Computer Vision*.

Zhihui Li, Lina Yao, Xiaoqin Zhang, Xianzhi Wang, Salil Kanhere, and Huaxiang Zhang. 2019. Zero-shot object detection with textual descriptions. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

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

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019a. Linguistic knowledge and transferability of contextual representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings of the 33rd Conference on Advances in Neural Information Processing Systems*.

Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2020. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the 2020 IEEE Conference on Computer Vision and Pattern Recognition*.

Li Lucy and Jon Gauthier. 2017. 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, Vancouver, Canada. Association for Computational Linguistics.

Qiaomei Mao, Chong Wang, Shenghao Yu, Ye Zheng, and Yuqi Li. 2020. Zero-shot object detection with attributes based category similarity. *IEEE Transactions on Circuits and Systems II: Express Briefs*.

James L McClelland, Felix Hill, Maja Rudolph, Jason Baldridge, and Hinrich Schütze. 2019. Extending machine language models toward human-level language understanding. *arXiv preprint arXiv:1912.05877*.

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

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the Empirical Methods in Natural Language Processing (EMNLP)*.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv e-prints*.

Shafin Rahman, Salman Khan, and Fatih Porikli. 2018. Zero-shot object detection: Learning to simultaneously recognize and localize novel concepts. In *Proceedings of the Asian Conference on Computer Vision*. Springer.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *arXiv preprint arXiv:2002.12327*.

Mohammad Amin Sadeghi and Ali Farhadi. 2011. Recognition using visual phrases. In *CVPR 2011*, pages 1745–1752. IEEE.

John R Searle, S Willis, et al. 1984. *Minds, brains, and science*. Harvard University Press.

Xing Shi, Inkit Padhi, and Kevin Knight. 2016. Does string-based neural MT learn source syntax? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Carina Silberer and Mirella Lapata. 2014. Learning grounded meaning representations with autoencoders. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. 2013. Zero-shot learning through cross-modal transfer. In *Advances in Neural Information Processing Systems*.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. Vl-bert: Pre-training of generic visual-linguistic representations. In *Proceedings of the 2020 International Conference on Learning Representations*.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. 2019. olmpics–on what language model pre-training captures. *arXiv preprint arXiv:1912.13283*.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

Hao Tan and Mohit Bansal. 2020. Vokenization: Improving language understanding with contextualized, visual-grounded supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*.

Pengkai Zhu, Hanxiao Wang, and Venkatesh Saligrama. 2019. Zero shot detection. *IEEE Transactions on Circuits and Systems for Video Technology*.