Probing 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

While large-scale language models have enjoyed great success recently, much remains to be understood about what is encoded in their representations. In this work, we propose a method for characterizing how language representations of concrete nouns relate to the physical appearance of the objects they refer to. Our approach uses a probing model that examines how useful language representations are in discerning between different visual representations. We show evidence of a surprising common ground with the visual domain, finding representations of many language models to be useful in retrieving semantically aligned image patches. In control experiments where language and visual representations are intentionally mismatched, we observe much weaker results. Furthermore, we examine the impact of textual context in our experiments, finding, for instance, that nouns accompanied by adjectives lead to more accurate retrieval. Finally, we show that the examined models substantially under-perform humans in retrieval. Altogether, our findings shed new empirical insights on language grounding, suggesting that some physical properties are being captured by trained language models, and highlighting large room for future progress.

1 Introduction

Recent years have been marked by substantial progress in learning representations from unlabelled text (Devlin et al., 2019; Liu et al., 2019b, among many others). Some argue that learning semantics from text only is reasonable under the distributional hypothesis (Harris, 1954)—the idea that the meaning of a word can be learned from the contexts it is used in. Others defend that grounding language by directly observing physical properties of concrete words is imperative for learning meaning (Harnad, 1990), and a key challenge for progress in language understanding (McClelland et al., 2019; Bisk et al., 2020; Bender and Koller, 2020). From a practical standpoint, we argue that empirically characterizing the relation between representations from text and vision models enables a better understanding of language grounding in recent literature.

In this work, we study neural representations of concrete concepts that can be visually observed, such as apple or banana. These representations are extracted using trained models, either from text or from images. As we will see in Section 2, language and visual representations learned independently exhibit local patterns that, despite being far from a perfect match, intersect significantly more than chance. In Sections 3-5, we expand on these results by inspecting language representations through the technique of probing (Figure 1).

In short, probing consists of training a supervised model, known as the probe, to predict certain properties from frozen representations (Shi et al., 2016; Conneau et al., 2018; Peters et al., 2018; Tenney et al., 2019). This approach is typically used in settings when discrete linguistic annotations such
as parts of speech are available (Rogers et al., 2020). Here, using a contrastive loss (Oord et al., 2018), we probe language models for common ground with visual representations. In training, the probe learns to map language representations of concrete objects in captions to visual representations from a semantically aligned image patch. Once trained, the probe is evaluated by retrieving image patches from its outputs, using object categories either seen or unseen by the probe during training.

We examine representations from multiple language models, including ones trained on purely textual data—GloVe, BERT, RoBERTa and ALBERT (Pennington et al., 2014; Devlin et al., 2019; Liu et al., 2019b; Lan et al., 2019)—and on vision and language—LXMERT, VL-BERT and VILBERT-MT (Tan and Bansal, 2019; Su et al., 2020; Lu et al., 2019a,b). Following common practice in recent language grounding literature, we use visual features from Faster R-CNN (Ren et al., 2015) trained on Visual Genome (Krishna et al., 2017).

For all these models, we are able to learn non-trivial mappings from language to visual representations. For instance, visual patches with correct, unseen object categories are retrieved 84-91% of the time using data from MS-COCO. Compared to models trained only on text, we find similar, but generally more accurate results for vision and language models. Further, we show that contextual models, that use the entire sentence when building representations, substantially outperform non-contextual GloVe embeddings. By comparing recall from sentences where objects are or are not accompanied by adjectives, we provide further insights of how context affects retrieval performance. 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, particularly for unseen object categories.

Finally, we turn to human judgment to assess an upper-bound in performance. The examined models significantly under-perform humans in mapping language to visual inputs. For instance, when compared to humans, BERT shows a 33% absolute decrease in recall performance at the instance level, exposing much room for improvement in representation learning and natural language grounding.

Our main findings and contributions are to: 1) Introduce a probing procedure for examining the relation between language and visual representations; 2) For multiple models, find that language representations are useful in choosing between different visual representations, a much weaker result in control experiments; 3) Examine how our experiments are affected by context when building language representations; 4) Assess human performance in retrieval, suggesting large room for future progress.

2 Language and visual representations

At the center of our analysis are continuous representations of concrete concepts. More specifically, we are interested in nouns that can be observed visually, which we refer to as object categories. Our inquiry relies on pairs \((\ell, v)\) of semantically aligned instances of text \(\ell\) and visual \(v\) representations of these object categories, which we collect from trained models and image captioning data, as formalized below.

**Language representations** \((\ell)\) are extracted from image captions. To accommodate recent natural language processing models and tokenizers, we allow such representations to be contextual\(^1\) and have variable length\(^2\), where each element \(\ell_i \in \ell\) has a fixed dimension \(d_{\ell}\). The length of the representations \(\ell\) for each object category is determined by the tokenizer. We treat a model that extract representations from text as a function \(\Lambda\) 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 vision and language models, when extracting representations from text, and 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 \(\Theta\), namely Faster R-CNN. 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 object category \(o\) and 2) assigned the highest confidence by the model among those. Visual representations \(\Theta(o \mid i)\) have fixed dimensions \(d_v\).

---

\(^1\)This allows models to fully use the caption when building representations, permitting, for instance, disambiguation of *bat* in a boy hits the ball with the bat, which here does not refer to the animal.

\(^2\)Conforming with sub-word tokenizers or multi-word expressions such as *fire extinguisher*, we let \(\ell = (\ell_1, \ldots, \ell_n)\).
**Paired data** $(\ell, v)$ with semantically 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 $\Theta$, 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 $(\ell, v)$ associated with the same visual instance, we ensure that at most one pair $(\ell, v)$ per object category in each image is included. In this work, paired representations are collected from the MS-COCO 2015 Image Captioning Task (Lin et al., 2014), with over 120 thousand images and 600 thousand captions. We use the 1600 object categories from Faster R-CNN (Ren et al., 2015) trained on Visual Genome (Krishna et al., 2017).

**Inspecting local patterns.** We briefly turn our attention to object categories tokenized as a single token for BERT base and VILBERT-MT, which correspond to the majority of categories studied in this work. We find that some common local structure emerge between these fixed-dimensional language representations and visual representations from Faster R-CNN, as illustrated in Figure 2. We compute the average cosine distance between pairs of representations of every two object categories. For a given object category, this measure allows ranking all other object categories. Analyses of the most similar object categories according to this measure can be found in Tables 1 and 2, showing further evidence of similarities between representations in both domains. While these analyses are limited in scope, examining only category-level, local patterns for object categories with a single token, we define our more general probing procedure in the next section.

### 3 Probing representations

At a high level (Figure 3), our approach uses a shallow neural model that maps language to visual representations. This probe is driven to output features able to distinguish the correct visual representations from a set of distractors. Once trained, it can map arbitrary language representations to the visual space, and evaluated by computing its recall in retrieving semantically aligned image patches.

#### 3.1 Training procedure

Neural representations containing either language or visual features can be thought of as random variables conditioned on raw inputs (text or images). In this work, the probe is optimized to maximally preserve the mutual information between these distributions, through the InfoNCE loss (Oord et al., 2018). Formally, from pairs of semantically aligned representations $(\ell, v)$, a probing model $\Psi_\theta$ learns parameters $\theta$ while minimizing the loss in Equation 1.

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

The probe $\Psi_\theta$ takes inputs $\ell$ and estimates visual representations $\hat{v} = \Psi_\theta(\ell)$ with the same dimen-

<table>
<thead>
<tr>
<th>Base Model</th>
<th>Most similar obj. categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>shorts Bb</td>
<td>jeans, pants, trunks, skirt</td>
</tr>
<tr>
<td>FR</td>
<td>skirt, pants, trunks, jeans</td>
</tr>
<tr>
<td>newspaper Bb</td>
<td>magazine, book, radio, paper</td>
</tr>
<tr>
<td>FR</td>
<td>magazine, papers, receipt, menu</td>
</tr>
</tbody>
</table>

Table 1: For each object category, we rank other categories by the cross-category average cosine similarity between instances in each. The table shows the average cross-modal intersection over union of the top-k neighbors of object categories, ranked using different language representations and Faster R-CNN.

Table 2: Most similar object categories to shorts and newspaper, for BERT base (Bb) and Faster R-CNN (FR), computed using average cosine distance between representations.
sionality \( d_V \) as the corresponding visual representations \( v \). For each pair \((\ell, v)\), this loss relies on a set of distractors \( V_i^{\text{NEG}} \), containing visual representations not positively associated with the language representations \( v \). The representations in \( V_i^{\text{NEG}} \) are used for contrastive learning and can be 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 \( u \in V_i^{\text{NEG}} \). In other words, training pushes the estimates \( \hat{v} = \Psi_\theta(\ell) \) to be maximally useful in discerning between positive and negative visual pairings.

In practice, the expectation in Equation 1 is estimated over a batch of size \( B \) with samples of language \( L = (L_1, \ldots, L_B) \) and visual representations \( V = (V_1, \ldots, V_B) \), where representations with the same index are associated. For efficiency, we use all other visual representations in the batch as distractors for a given representation \( V_i^{\text{NEG}} = \{V_j, j \neq i\} \). Thus, only the pairwise dot products \( \langle \hat{V}_i = \Psi_\theta(L_i), V_j \rangle \) are needed to calculate the loss, as illustrated in Figure 3. The batch loss function is then written as:

\[
\mathcal{L}_B = -\frac{1}{B} \sum_{1 \leq i \leq B} \left[ \log \frac{\exp(\langle \hat{V}_i, V_i \rangle)}{\sum_{1 \leq j \leq B} \exp(\langle \hat{V}_i, V_j \rangle)} \right]
\]  

Equation 2 is the final loss used for training the parameters \( \theta \) of the probe. Importantly, we note that the models used to extract representations are not trained or changed in any way during the probing procedure.

### 3.2 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. Let \( \mathcal{V} \) represent the set of all collected visual representations for evaluation. For each language representation \( \ell \), 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 \( k \) under two scenarios:

- **Instance Recall (IR@k)** is the percentage of pairs \((\ell, v)\) where the instance \( v \) is in the top \( k \) visual representations retrieved from \( \hat{v} = \Psi_\theta(\ell) \).

- **Category Recall (CR@k)** is the percentage of pairs \((\ell, v = \Theta(o \mid i))\) where any of the top \( k \) retrieved visual representations \( v' = \Theta(o' \mid i') \) is associated with the same object category \( o \) as \( v \) (i.e. \( o' = o \)).

To avoid conclusions over specific sets of object categories, we evaluated our probe in two scenarios: where pairs \((\ell, v)\) were collected using object categories either seen or unseen by the probe during training. For both scenarios, images and captions have no intersection with those used in training. Finally, we create multiple seen/unseen splits from
our data, training and testing on each split. We then report average and standard deviation for recall scores.

4 Experimental settings

4.1 Language models

We examine representations from multiple models, trained using purely textual inputs or multi-modal data. When applicable, we use representations extracted by the last layer of the models.

Text-only. 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 and ALBERT (Devlin et al., 2019; Liu et al., 2019b; Lan et al., 2019). For all these models, we use pre-trained weights from the HuggingFace Transformers library (Wolf et al., 2019). 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.

Vision and language. We examine multiple models trained on vision and language tasks, namely LXMERT, VL-BERT (base and large) and VILBERT-MT (Tan and Bansal, 2019; Su et al., 2020; Lu et al., 2019a,b)). These are transformer-based models based on self-attention. When necessary, we adapt them to include only the language branches, restricting attention to the text inputs. For all models, we use the code and weights made public by the authors.

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., 2019b), we use a Faster-RCNN 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)\(^6\), and do not fine-tune during probe training.

4.3 Data

We build disjoint training, validation and test sets from the aggregated training and validation sets of MS-COCO. To examine how our findings generalize to new objects, we test on representations from either seen or unseen object categories, built from images and captions not present in the training data. We use 1400 out of the 1600 object categories for training and seen evaluation, and reserve the remaining 200 for unseen evaluation. Furthermore, we perform train and test our probe multiple times, each with a different 1400/200 split of the object categories. While their exact size varies with specific object categories splits, each training set contains at least 300 thousand pairs of representations. For consistency, the seen test set is truncated to 7000 data pairs, and the unseen test set to 1000 pairs.

4.4 Control task

Central to probing is the practice of contrasting performance with a control task (Hewitt and Liang, 2019). We follow this practice by learning from a control task where representations are mapped permuted visual representations. More precisely, we replace each visual representation $v = \Theta(o | i)$ with another $v' = \Theta(o' | i')$ from an object category $o' = f(o)$ that depends on the original object category $o$. For instance, all visual representations with the original category cat are replaced with representations from a second category dog; visual representations from the object category dog are replaced by those from tree, and so on.

4.5 Implementation details and hyper-parameters

Our probe consists of a shallow neural model. To process the naturally sequential language representations $\ell$, 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 representation space. The probe is trained using Adam optimizer (Kingma and Ba, 2014) with a learning rate of 0.0005, weight decay of 0.0005 and default remaining coefficients ($\beta_1 = 0.9$ $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$). We train with a batch size of 3072, for a total of 5 epochs on one GPU.
Table 3: Average instance recall (IR@k) and category recall (CR@k) for test sets with seen and unseen object categories. For a wide range of models, nontrivial mappings to visual representations can be learned. Unlike the control task with permuted representations, these mappings generalize well to unseen object categories. For each model, we train and evaluate 5 times, using different sets of object categories seen in training.

5 Results and discussion

For all examined language models, nontrivial and generalizable mappings to visual representations are learned. As shown in Table 3, recall scores are significantly better than random for all studied models (Rows 2-12). Moreover, the learned mappings generalize well to the test set with unseen object categories. Interestingly, we note from Table 3 that there is no strong correlation between retrieval performance and performance on purely textual tasks (for instance, at SQUAD (Rajpurkar et al., 2018) or GLUE (Wang et al., 2019)). This reinforces the intuition that text-only benchmarks are not the ideal landscape for studying language grounding.

Probe selectivity. Results from the control task with permuted representations are shown in Row 1 of Table 3, using representations from BERT. In this setting, we find a substantially lower performance compared to the base experiments, a gap specially high for unseen object categories. This attests to the high selectivity of the probe, which only yields good performance when representations were sensibly paired. We find similar results for other models, and refer to the Appendix for specific numbers.

Vision and language models. As shown in Table 3, representations from models trained on vision and language (Rows 9-12) also allow nontrivial mappings to be learned. Compared to models trained on text only (Rows 2-8), we find similar, but generally better performance. Larger differences are observed in instance retrieval, where representations need to be more comprehensive for mappings to be successful.

Influence of context. Contrasting the performance of non-contextual representations from GloVe with the remaining contextual models shows that context considerably affects instance recall. For instance, for unseen object categories, using representations from GloVe and BERT base leads to 5.1% to 12.0% IR@1, respectively. This gap is not surprising, since a non-contextual representation of an object category should not be able to discern between distinct image patches depicting it. We observe higher category recall for GloVe for seen object categories, which we hypothesize is due to the increased facility in correctly predicting the
Figure 5: Qualitative examples of visual patches (orange boxes in the right) retrieved from contextual representations of objects in captions (orange words in the left) extracted by BERT base. All shown samples are from images and captions from MS-COCO with object categories previously unseen by the probe. On the bottom rows, we show an example of how context can affect instance recall.

Figure 6: Instance recall at 1 when object categories are or are not accompanied by adjectives. For both BERT base and VILBERT-MT, more descriptive contexts led to more accurate retrieval.

We further investigate the influence of context by measuring performance of contextual models when the objects have 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). As shown in Figure 6, we observe conspicuous gains in instance recall for contextual models when objects are accompanied by adjectives.

Qualitative results. Figure 5 shows qualitative examples of retrieval using representations from BERT base, for unseen object categories. The bottom rows highlight an example where more descriptive sentences lead to better instance recall. For these examples, contextual representations are extracted from the same word cat, yet the context encoded in them allows the retrieval of better matching image patches.

Human performance. Mappings from all examined models substantially under-perform humans. Similar to probe evaluation, we measure human performance in retrieving visual patches from words in sentences. In virtue of the limited human attention, we evaluate on a reduced test set of 100 samples of unseen object categories, asking subjects to choose out of 100 image patches the closest match to an object in a sentence. We collect over 1000 annotations from 17 subjects, with at least 30 annotations each.
Table 4: A sizable gap in instance recall is seen by comparing the performance of humans and the examined models in a smaller test set with 100 samples.

<table>
<thead>
<tr>
<th>IR@1</th>
<th>Human</th>
<th>BERT base</th>
<th>VILBERT-MT</th>
</tr>
</thead>
<tbody>
<tr>
<td>76%</td>
<td>43%</td>
<td>53%</td>
<td></td>
</tr>
</tbody>
</table>

Our results (Table 4) show a large gap to mappings learned from representations extracted by the examined models: while humans retrieved the correct instance (IR@1) 76% of the time on average, the best performing model trained on text-only, BERT base, retrieved the correct instance only 43%, and the overall best model, VILBERT-MT, achieved an IR@1 of 53%.

### 6 Related Work

**What is encoded in language representations.** Understanding what information state-of-the-art NLP models encode has gained increasing 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 in this literature is the use of probes (Shi et al., 2016; Adi et al., 2016; Conneau et al., 2018; Hewitt and Liang, 2019), supervised models trained on top of frozen representations in some specific task of interest. Such models are typically used in settings were discrete, linguistic annotations are available. Our approach differs from previous work in both scope and methodology, focusing on probing language representations for similarities with continuous, visual representations.

**Zero-shot detection.** Recent work attempts to build object detectors that generalize to previously unseen object categories, through 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, we differ from this line of work in motivation, which translates to further experimental differences: given our goal to analyse the commonalities between already trained language models and vision models (as opposed to learning a generalizable object detector), we train nothing apart from a lightweight probe in our analyses.

**Language grounding.** A widely investigated research direction aims to connect natural language to the physical world (Bisk et al., 2020; McClelland et al., 2019), typically through training and evaluating models in multi-modal tasks (Antol et al., 2015; Hudson and Manning, 2019; Suhr et al., 2018; Zellers et al., 2019, among others). Closer to our motivations of understanding what is encoded in already trained language models are the works of (Lucy and Gauthier, 2017) and (Scialom et al., 2020). Lucy and Gauthier (2017) evaluate how representations from non-contextual word embeddings can predict discrete, human-generated perceptual features. Our work focuses instead on examining the semantic overlap with continuous visual representations, as they bear a direct, instance-specific connection to raw visual inputs. Scialom et al. (2020) examines cross-modal transferability of language models when using multi-modal inputs for text generation, finding that semantic abstractions from BERT generalize well to the visual domain. While our conclusions are generally aligned, our methodology differs significantly—instead of encoding visual representations into language models, we map purely textual representations to the visual domain.

### 7 Conclusion

In this work we proposed a procedure for evaluating similarities between latent representations extracted by language and by vision models. We find the representations of a wide range of language models to be useful in discerning between different visual representations. Moreover, these findings generalize well to unseen object categories, unlike the case with control experiments. These results do not imply that language models and vision models learn representations with perfect (or close to perfect) similarity, or that language grounding is trivial because of the commonalities that do exist. Our intention is to better understand the landscape of representation learning and natural language grounding empirically, and in so, encourage further progress in the fields. As suggested by human retrieval performance, there remains an appreciable headroom for building better grounded representations.
References


Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R Bowman, Dipanjan Das, et al. 2019. What do you learn from
probing for sentence structure in con-


### Appendix

<table>
<thead>
<tr>
<th>Text repr.</th>
<th>IR@1 Seen</th>
<th>Unseen</th>
<th>IR@5 Seen</th>
<th>Unseen</th>
<th>CR@1 Seen</th>
<th>Unseen</th>
</tr>
</thead>
<tbody>
<tr>
<td>GloVe</td>
<td>1.9 ± 0.1</td>
<td>0.1 ± 0.1</td>
<td>8.6 ± 0.5</td>
<td>0.6 ± 0.3</td>
<td>43.6 ± 4.5</td>
<td>1.6 ± 2.5</td>
</tr>
<tr>
<td>BERT base</td>
<td>1.6 ± 0.1</td>
<td>0.0 ± 0.0</td>
<td>7.8 ± 0.6</td>
<td>0.4 ± 0.2</td>
<td>41.3 ± 5.6</td>
<td>3.0 ± 1.4</td>
</tr>
<tr>
<td>BERT large</td>
<td>1.6 ± 0.1</td>
<td>0.1 ± 0.1</td>
<td>7.4 ± 0.3</td>
<td>0.3 ± 0.1</td>
<td>37.0 ± 4.2</td>
<td>1.9 ± 0.8</td>
</tr>
<tr>
<td>RoBERTa base</td>
<td>1.6 ± 0.2</td>
<td>0.1 ± 0.1</td>
<td>7.6 ± 0.4</td>
<td>0.4 ± 0.2</td>
<td>41.6 ± 3.6</td>
<td>4.0 ± 5.3</td>
</tr>
<tr>
<td>RoBERTa large</td>
<td>1.4 ± 0.2</td>
<td>0.0 ± 0.0</td>
<td>7.0 ± 0.6</td>
<td>0.6 ± 0.3</td>
<td>40.9 ± 5.4</td>
<td>2.7 ± 1.7</td>
</tr>
<tr>
<td>ALBERT base</td>
<td>1.5 ± 0.1</td>
<td>0.3 ± 0.1</td>
<td>7.3 ± 0.4</td>
<td>0.9 ± 0.2</td>
<td>35.9 ± 3.8</td>
<td>6.9 ± 5.0</td>
</tr>
<tr>
<td>ALBERT large</td>
<td>1.5 ± 0.1</td>
<td>0.1 ± 0.0</td>
<td>7.3 ± 0.5</td>
<td>0.6 ± 0.3</td>
<td>39.3 ± 4.8</td>
<td>2.2 ± 1.4</td>
</tr>
<tr>
<td>LXMERT</td>
<td>1.6 ± 0.2</td>
<td>0.1 ± 0.1</td>
<td>7.8 ± 0.7</td>
<td>0.4 ± 0.2</td>
<td>41.8 ± 6.5</td>
<td>1.2 ± 0.6</td>
</tr>
<tr>
<td>VL-BERT base</td>
<td>1.5 ± 0.1</td>
<td>0.1 ± 0.1</td>
<td>7.2 ± 0.3</td>
<td>0.4 ± 0.1</td>
<td>35.7 ± 4.0</td>
<td>1.8 ± 2.2</td>
</tr>
<tr>
<td>VL-BERT large</td>
<td>1.6 ± 0.2</td>
<td>0.1 ± 0.1</td>
<td>7.7 ± 0.5</td>
<td>0.4 ± 0.3</td>
<td>38.4 ± 5.6</td>
<td>1.6 ± 1.2</td>
</tr>
<tr>
<td>VILBERT-MT</td>
<td>1.6 ± 0.2</td>
<td>0.1 ± 0.1</td>
<td>7.7 ± 0.7</td>
<td>0.6 ± 0.4</td>
<td>39.7 ± 6.2</td>
<td>1.6 ± 0.4</td>
</tr>
</tbody>
</table>

Table 5: Average instance recall (IR@k) and category recall (CR@k) for z control task with permuted representations, in test sets with seen and unseen object categories. We see poor retrieval performance, especially for unseen object categories, indicating that the probe does not generalize satisfactorily. For each model, we train and evaluate 5 times, using different sets of object categories seen in training.