Supplementary Material
Investigating the Role of Attribute Context in Vision-Language Models for Object Recognition and Detection

Kyle Buettner[1], Adriana Kovashka[1,2]
[1]Intelligent Systems Program, [2]Department of Computer Science, University of Pittsburgh, PA, USA
buettnerk@pitt.edu, kovashka@cs.pitt.edu

1. Supplementary Material

The supplementary material contains two sections. Section 1.1 outlines how we generate data for training and evaluation. Section 1.2 shows further experiments for phrase grounding analysis of OVR-CNN.

1.1. Data Generation

1.1.1 Classification via description

We present details regarding data generation for measurement of CLIP's attribute sensitivity, particularly for use in the classification by description task on ImageNet-v2 [7]. Overall, three "styles" of CLIP prompts are used in inference: (1) CLIP's default "a photo of" prompts, (2) LLM-based *multiple descriptor* prompts, and (3) LLM-based *single sentence* prompts.

To produce (1), all class names from ImageNet-v2 are filled into the following template and processed by CLIP's text encoder to produce classifier weights:

> "a photo of a <category>."

For instance, CLIP's classifier would contain the text encodings of "a photo of a petri dish", "a photo of a basketball", etc. (all 1,000 classes).

For (2), we use the methodology of Menon and Vondrick [5] to produce descriptions with attribute context. For every class in ImageNet-v2, we prompt GPT-3 (*davinci-002*, max token length 100, temperature 0.7) with a *multiple descriptor* prompt template:

> Q: What are useful features for distinguishing a lemur in a photo?
> A: There are several useful visual features to tell there is a lemur in a photo:
> - four-limbed primate
> - black, grey, white, brown, or red-brown
> - wet and hairless nose with curved nostrils
> - long tail
> - large eyes
> - furry bodies
> - clawed hands and feet
> Q: What are useful features for distinguishing <category> in a photo?
> A: There are several useful visual features to tell there is/are <category> in a photo:

Output descriptors are returned in the format of the lemur example. We further process these outputs with the following CLIP prompt template:

> <category> which (is/has/etc) <descriptor>.

There are thus multiple prompts for each class. There would be seven prompts for the lemur example, for instance. These would appear as:

> - a lemur which (is/has/etc) a four-limbed primate.
> - a lemur which (is/has/etc) black, grey, white, brown, or red-brown.
> - ...

The average CLIP similarity between the image of interest and each descriptor (Eq. 7 in main paper) is used as the score for each class in classification.

For (3), we use a *single-sentence* description-based prompt template for GPT-3, in particular the one from [6]:

> Q: What does a lorikeet look like? Describe with one sentence.
> A: A lorikeet is a small to medium-sized parrot with a brightly colored plumage.
> Q: What does <category> look like? Describe with one sentence.
> A:

We directly use each class's result from GPT-3 as a respective CLIP prompt. For instance, the lorikeet's CLIP prompt would be:

A lorikeet is a small to medium-sized parrot with a brightly colored plumage.

The lemur's CLIP prompt could be:

A lemur is a small, four-limbed primate with large eyes, a long tail, and a slender body.

After generating the prompts for (1)-(3), we remove/change adjectives detected with spaCy [2] (v3.5.3) in all class prompts for inference. For class name removal, we replace *all* class names with "a/an object" when creating the CLIP classifier. As an example, consider possible prompts used to create CLIP's class embeddings with (3):

(P1) A baseball is a round, stitched ball made of leather or synthetic materials, typically white with red stitching.
(P2) A hockey puck is a flat, disk-shaped object made of hard rubber, often black in color, used in the sport of ice hockey.
...
(P1000) A basketball is a round, inflatable ball with a synthetic or leather cover, black lines, and typically orange in color.

Removed class names would change the classifier to:

(P1) An object which is a round, stitched ball made of leather or synthetic materials, typically white with red stitching.
(P2) An object which is a flat, disk-shaped object made of hard rubber, often black in color, used in the sport of ice hockey.
...
(P1000) An object which is a round, inflatable ball with a synthetic or leather cover, black lines, and typically orange in color.

In Table 1, we provide key statistics regarding the descriptors generated for this analysis.

Statistic	CWD [5]	SS [6]
Avg. # Descriptions Per Class	5.29	1
Avg. Desc. Length (spaCy tokens)	18.43	20.65
Total # Adjectives Perturbed	4,831	2,392
# Unique Adjectives	607	501
# Adjectives Per Description	0.91	2.39

Table 1. Statistics for measuring attribute sensitivity in classification via description. CWD=Classification With (Multiple) Descriptors; SS=Single Sentence. Note that there are more adjectives/description in the single-sentence case, potentially explaining its increased sensitivity to "Change ADJ" in Fig. 5 (main paper).

1.1.2 Analyzing COCO

Given the role of COCO [1, 3] in pretraining OVR-CNN, finetuning OVR-CNN/CLIP, and gauging the attribute sensitivity of OVR-CNN, we provide specific details regarding its usage. We describe in detail how we identify objects of interest and attribute context belonging to those objects, as well as statistics related to the data creation process.

For all of these tasks, we create a vocabulary $\mathcal{V}$ of terms/phrases corresponding to COCO class names (*e.g.* "car", "fire hydrant", "teddy bear"). We build $\mathcal{V}$ from the synonym list of COCO class names provided in [4], with plural terms also added. Adjectives used for specific COCO objects are detected using the spaCy dependency parser [2]. For each caption, we traverse the parse tree and mark all "amod" with dependency on a class term in $\mathcal{V}$, taking note of which class each term belongs to. We also mark terms not explicitly detected as "amod", but connected through coordinating conjunctions ("cc"). We use per-class lists to create the *plausible* sets. Some example top occurring plausible adjectives are shown in Table 2. All unique adjectives detected across all classes lie in the *random* set. We show the counts of unique adjectives detected across COCO classes in Fig. 1. These sets are used to sample negative caption adjectives in pretraining and to change plausibly/randomly in unsupervised phrase grounding.

COCO Class	Adjective	Count
bear	black	860
	brown	822
	polar	802
	large	486
	white	244
frisbee	white	302
	yellow	229
	red	201
	blue	132
	green	92
apple	green	182
	red	153
	sliced	38
	yellow	26
	several	22

Table 2. Examples of top 5 detected adjectives and counts for COCO categories. Note the frequent use of colors and state adjectives (*e.g. sliced, large*).

For OVR-CNN, COCO is used in our study in pretraining (2017train as an image-caption dataset), in finetuning (2017train/2017val for detection), and in unsupervised phrase grounding (2017val as an image-caption evaluation set). For CLIP, COCO is used in finetuning (2017train/2017val for image-text matching). COCO is

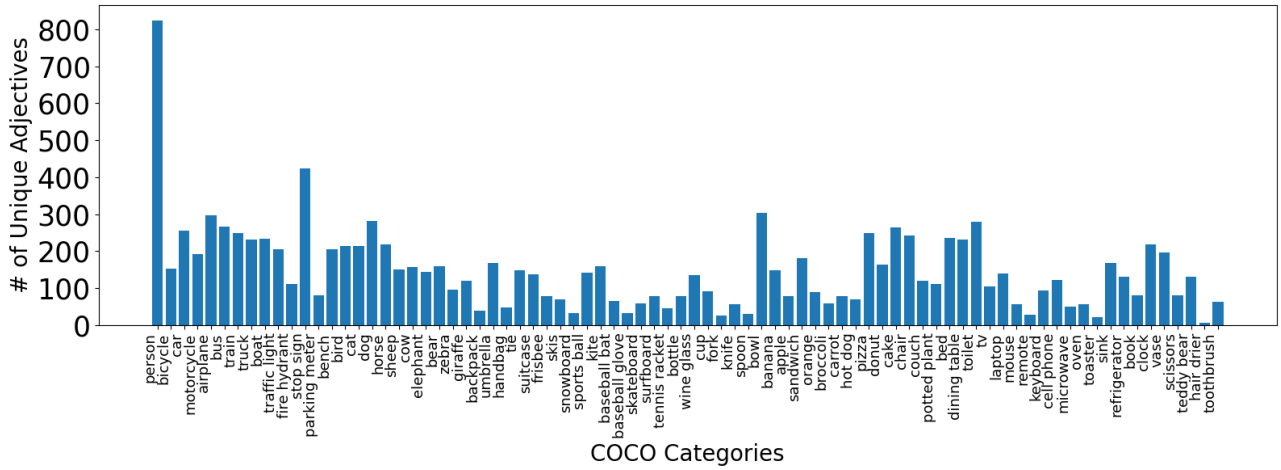

Figure 1. A histogram of unique adjectives described with objects in the COCO Captions corpus.

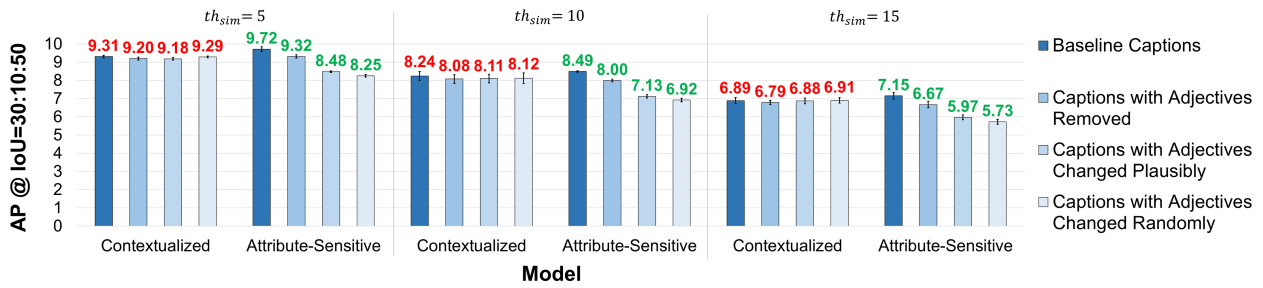

Figure 2. **Measuring attribute sensitivity in contextualized object grounding**. In an attribute-sensitive model, grounding performance should drop if an incorrect attribute is used, which occurs when changing adjectives. Through unsupervised grounding, we show that default contextualization does *not* result in substantial AP@IoU=30:10:50 drops vs. the baseline with adjectives changed (red), illustrating a lack of sensitivity to attribute meaning. When we add adjective negatives (plausible in this case), contextualization gains enhanced sensitivity to attribute meaning, shown in the decreases from baseline to changing (green). **Note that such trends hold over values of** th_{sim}. The presented values are averages over 3 pretraining trials, and error bars show standard error.

Statistic	COCOtrain	COCOval
# Total Captions	591,753	5,000
# Caps. with COCO AdjMod	153,207	1,294
Total # Adjectives Perturbed	191,772	1,611
# Unique Adjectives	3,080	277

Table 3. **Statistics for modifying adjectives in COCO captions (train/val).** AdjMod = token labeled as an adjectival modifier. Many captions (>150k) contain one or more adjectives, which represents a significant amount of signal that models can leverage.

used with both models for text-region retrieval (2017val). We provide further details of the detected adjectives in train/val in Table 3. For phrase grounding, we choose one caption for each image to use, resulting in 5,000 test cases.

1.2. Threshold experiments: Unsupervised phrase grounding

A hyperparameter for unsupervised phrase grounding is th_{sim}, which determines how bounding boxes are created from similarity maps. We find in practice that the attribute sensitivity trends of interest (i.e. relative AP@t differences between baseline/removal and baseline/changing) hold across values of this parameter for default contextualization and a model with adjective negatives. Fig. 2 shows three values we experiment with to support this finding.

References

[1] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 2

[2] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. SpaCy: Industrial-strength Natural Language

Processing in Python. 2020. 2

[3] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 2

[4] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Neural baby talk. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7219–7228, 2018. 2

[5] Sachit Menon and Carl Vondrick. Visual classification via description from large language models. *International Conference on Learning Representations, ICLR*, 2023. 1, 2

[6] Sarah Pratt, Rosanne Liu, and Ali Farhadi. What does a platypus look like? Generating customized prompts for zero-shot image classification. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 1, 2

[7] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *International Conference on Machine Learning*, pages 5389–5400. PMLR, 2019. 1