Supplementary Material: Improving Zero-shot Generalization and Robustness of Multi-modal Models

Yunhao Ge[1,2*], Jie Ren[1*], Andrew Gallagher[1], Yuxiao Wang[1], Ming-Hsuan Yang[1], Hartwig Adam[1], Laurent Itti[2], Balaji Lakshminarayanan[1†], Jiaping Zhao[1†]
[1]Google Research [2]University of Southern California
[*]co-first author, [†]correspondence to {balajiln, jiapingz}@google.com

Appendix

A. Analyzing the classes for which the top-5 prediction is correct but the top-1 prediction is mostly incorrect

Ground Truth Class Name	Error rate
tusker	94%
missile	94%
terrapin	92%
collie	90%
screw	90%
mushroom	88%
Appenzeller Sennenhund	84%
snoek fish	84%
husky	82%
parallel bars	82%
gazelle	82%
sailboat	82%
corn cob	80%
analog clock	78%
cornet	78%
gossamer-winged butterfly	76%
green mamba	76%
tiger cat	74%
hare	74%
canoe	72%

Table 1. List of top 20 classes where the top-5 prediction is correct but the top-1 prediction is mostly incorrect: sorted descendingly by error rate. The % indicates the proportion of images within the class whose top-5 prediction is correct but whose top-1 prediction is incorrect.

B. Locating the 1000 ImageNet classes at WordNet hierarchy

Fig. 1 shows the location of the 1000 ImageNet classes within the WordNet hierarchy. The 1000 ImageNet class names are at different levels of WordNet hierarchy with different degrees of abstraction. 350 are super-classes with sub-classes as the children, while the remaining 650 are leaf nodes with no children. (b) The distribution of the number of children: 12.7% of the classes have one child node and 16.6% have 2-4 child nodes.

C. Additional results

Qualitative visualization. Figure 2 shows a qualitative visualization on more typical failure modes in the cases where our top-down and bottom-up prompt augmentation using the WordNet hierarchy method fixes the error.

Effect of the model architecture and size on selected low-confidence sets on ImageNet. We found a more powerful backbone leads to a smaller low-confidence sets (e.g., the low-confidence sets of ViT-l14 and ResNet-50 contain 8,557 and 13,106 images, respectively).

Benefits to Top-5 accuracy. If we apply our top-down and bottom-up label augmentation method to re-rank top-10 classes, we see it can improve the top-5 on the low confidence set from 77.4% to 80.2%. We also find reranking top-10 further improves top-1 performance vs. re-ranking top-5 only.

Sparsifying WordNet using the norm of text embedding. WordNet contains many academic words that are rarely used in common usage of English, and hence unlikely to occur frequently in the captions used for CLIP training. For example, "anthozoan, actinozoan", "coelenterate", "gastropod", etc.. Directly using the raw WordNet with academic words as parents is not helpful for improving zero-shot accuracy, and can even hurt the performance. Though we do not have access to the CLIP private data, we studied the norm of the word embedding vector and found it is correlated with word frequency. We compute the L_2 norm of the prompt embedding when plugging in the word into promt

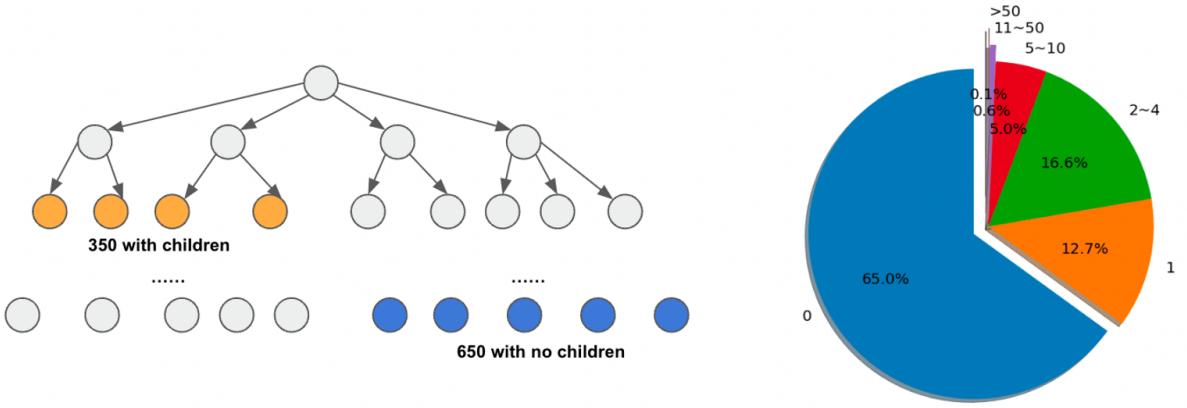

Figure 1. (a) The 1000 ImageNet class names are at different levels of WordNet hierarchy with different degree of abstraction. 350 of them are super-class with sub-classes as the children, while the rest 650 of them have no children. (b) The distribution of the number of children: 12.7% of the classes have one child node, 16.6% of the classes have 2-4 child nodes.

Ground Truth: Collie

Misclassified as: Shetland Sheepdog

Parent: Dog

90% of images with ground truth label "collie" are wrongly classified as other dog classes such as "Shetland Sheepdog". Concatenating the parent class name "dog" fixes such errors.

Ground Truth: mink

Misclassified as: European polecat

Child: American mink

78% of images with ground truth label "mink" are wrongly classified as other animal classes such as "European polecat". Using child class names for "mink" (e.g. "American mink") fixes such errors.

Ground Truth: motorboat

Misclassified as: trimaran

Child: hydrofoil

62% of images with ground truth label "motorboat" are wrongly classified as other boat classes such as "trimaran". Using child class names for "motorboat" (e.g. "hydrofoil") fixes such errors.

Figure 2. Qualitative visualization on typical failure modes for cases where our top-down and bottom-up prompt augmentation using the WordNet hierarchy method fixes the errors. In each case, the image is originally mis-classified but is correctly classified with our proposed method.

templates, i.e., $\|f_{text}(t(c))\|, t \in \mathcal{T}$. We found that the variance of the norm, $\text{Var}_{t\in\mathcal{T}}(\|f_{text}(t(c))\|)$, is correlated with word frequency. Rare words tend to have small variances, while common words tend to have large variances. For example, the variance of the rare word "anthozoan" is 0.118, while the variance of a more common word "workplace" is 0.724. We use this statistic to filter out rare words in WordNet. We removed 60% of the nodes in WordNet and only kept the top 40% nodes with the highest variance and found this may work slightly better than using the whole WordNet in some cases. Our intuition behind the correlation between the norm variance and the word frequency is that, for a frequent word that has many examples in the CLIP training data, the CLIP model learns a very precise text embedding such that it has the capability to tell the semantic difference under different contexts, e.g., "`a photo of a nice {label}`" and "`a photo of a weird {label}`".

Table 2. Effect of WordNet sparsity on zero-shot top-1 accuracy on ImageNet with CLIP.

% of remaining words	acc overall
100%	68.52%
40%	68.52%
30%	**68.72%**
20%	**68.72%**
10%	**68.72%**

Effect of WordNet sparsity on zero-shot accuracy We evaluate the effect of the degree of sparsity of WordNet on the downstream zero-shot accuracy. We sparsify the WordNet based on word frequency, which is measured by embedding variance as described in the previous section. Here we study the effect of sparsity on the downstream zero-shot accuracy. Table 2 shows the overall accuracy on ImageNet using CLIP model with different levels of WordNet sparsities .