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

Abstract

Multi-modal image-text models such as CLIP and LiT have demonstrated impressive performance on image classification benchmarks and their zero-shot generalization ability is particularly exciting. While the top-5 zero-shot accuracies of these models are very high, the top-1 accuracies are much lower (over 25% gap in some cases). We investigate the reasons for this performance gap and find that many of the failure cases are caused by ambiguity in the text prompts. First, we develop a simple and efficient zero-shot post-hoc method to identify images whose top-1 prediction is likely to be incorrect, by measuring consistency of the predictions w.r.t. multiple prompts and image transformations. We show that our procedure better predicts mistakes, outperforming the popular max logit baseline on selective prediction tasks. Next, we propose a simple and efficient way to improve accuracy on such uncertain images by making use of the WordNet hierarchy; specifically we augment the original class by incorporating its parent and children from the semantic label hierarchy, and plug the augmentation into text prompts. We conduct experiments on both CLIP and LiT models with five different ImageNet-based datasets. For CLIP, our method ***improves the top-1 accuracy by 17.13% on the uncertain subset and 3.6% on the entire ImageNet validation set.*** *We also show that our method improves across ImageNet shifted datasets, four other datasets, and other model architectures such as LiT.* ***The proposed method[1] is hyperparameter-free, requires no additional model training and can be easily scaled to other large multi-modal architectures.*** *Code is available at* https://github.com/gyhandy/Hierarchy-CLIP.

1. Introduction

Vision-language multi-modal models trained on large-scale data have achieved significant success in numerous domains and have demonstrated excellent zero-shot generalization ability [7, 12, 18, 19, 20, 28]. Given a test image and a set of candidate class labels, one can compute the similarity between the embedding of the image and the embedding of each candidate class labels, and predict the class as the one with the highest similarity. The zero-shot top-1 accuracy for ImageNet [4] using CLIP variants (CLIP ViT-L) matches the performance of the original ResNet model trained from scratch. Recently, CLIP has been found to be more robust to distribution shift than ResNet, achieving good performance on ImageNet-V2 [21], ImageNet-R [9], ImageNet-A [11], and ImageNet-Sketch [25].

We noticed a large gap between the top-1 accuracy and top-5 accuracy, 64.2% vs. 89.4% respectively, revealing potential headroom for improvement. We investigated the cases where the top-1 prediction was incorrect but the top-5 prediction was correct, and identified several typical failure modes. Despite the well-known multi-label issues in ImageNet [1], we found many of the remaining failure cases are caused by noise and ambiguous text prompts related to the WordNet hierarchical structure of ImageNet. Some class names are quite general so that the model cannot correctly match images from their specific subclasses. For example, the hot-air balloon images belonging to the "balloon" class were misclassified as "airship", see Figure 1 middle. On the other hand, some class names are too specific such that the model fails to correlate them with their more generic super-classes. For example, 96% of images with ground truth label "tusker" are wrongly classified as other elephant classes such as "Asian elephant", see Figure 1 left. The failure modes analysis suggests that the text encoder is very sensitive to inputs and as a result, the overall classification lacks robustness.

Inspired by these observations, we propose to first identify the subset of images whose top-1 prediction is likely to be incorrect, and then improve the accuracy for those images by a principled framework to augment their class labels by WordNet hierarchy. To estimate whether an image has an incorrect prediction, i.e., to estimate the prediction confidence, we use the consistency of predictions under different text prompt templates and image augmentations as a signal for prediction confidence estimation. Although prediction confidence estimation has been well studied in single-modal classification models, we found those commonly used confidence scores, maximum softmax proba-

[1]Work carried out mainly at Google

bility [10] and maximum logit score [8], are not always reliable for the multi-modal CLIP and LiT models due to the poor calibration of the logits scores. For example, among the 1K classes in ImageNet, the class with the greatest mean logit value (computed as the cosine similarity between image and text embeddings) is "fig" (the fruit). Though we don't have access to CLIP private training data, we hypothesize that this might be due to "fig" being a common abbreviation for "figure", which frequently occurs in the training data and thus includes many non-fruit illustrations.

In this work, we first propose a simple yet efficient zero-shot confidence estimation method better suited for CLIP, based on predictions' self-consistency over different text prompts and image perturbations. [26] proposed using self-consistency among multiple model outputs to improve the reasoning accuracy of large language models. Here we extend the idea for confidence estimation in multi-modal models by measuring *consistency of predictions under multiple input text prompts and image transformations*. Our method is effective at predicting mistakes; the identified low confidence subset has significantly lower top-1 accuracy (21.58%) than the average accuracy (64.18%). Next, to improve the accuracy for the low confidence subset, we develop a label augmentation technique using WordNet label hierarchy. Our method leverages semantic information from ancestors (top-down) as well as children (bottom-up) and improves the top-1 accuracy of the subset to 38.71% (17.13% improvement). Our method not only improves model accuracy, but also model robustness, improving on ImageNet variants with distribution shift such as ImageNet-v2, ImageNet-R, ImageNet-Adversarial and Imagenet-Sketch.

The main contributions of this work are:

- We identified several failure modes for zero-shot ImageNet classification using multi-modal models, and our findings suggest that the text encoder is very sensitive to prompts. To improve the prediction accuracy, prompts need to be better designed.
- We propose a simple yet efficient zero-shot confidence score that is better suited for multi-modal models, based on predictions' self-consistency under different text prompts and image perturbations.
- We develop a label augmentation technique that uses both ancestor and children labels from WordNet. By applying the label augmentation to the previously identified low confidence subset of images, we significantly improve their prediction accuracy.

2. Related work

Confidence estimation. Reliably estimating the confidence of a prediction is helpful for downstream decision making and can ensure the safe deployment of machine learning models. A well-calibrated confidence estimation should assign low scores for incorrect predictions and high score for correct predictions. Maximum softmax probability [10] and maximum logit [8] are the most commonly used confidence scores for classification problems, because of their simplicity and computational efficiency. Recent works propose more sophisticated confidence estimation methods which either involve modifications to the classification models or significantly increase the inference time. For example, Bayesian approaches such as Gaussian Process layer [16] and dropout-based variational inference [6] assume the weights in the neural networks are random variables such that the final prediction follows a distribution. A large variance of a prediction indicates the low confidence of the prediction. Non-Bayesian methods such as ensemble-based methods which aggregate the predictions from multiple models to improve the robustness of the confidence estimation [14, 27]. Those sophisticated methods were developed and studied in the single-modal models, and the application to multi-modal models is not straightforward. In addition, those methods mostly require modification to the model and additional training, which becomes challenging to multi-modal models since the training data are generally not publicly available. In our work, we focus on a zero-shot confidence estimation that is exclusively designed for multi-modal models. Our method does not require additional training, and is simple, efficient, and effective.

Prompt engineering. Prompt engineering and learning has attracted much attention in vision and learning since the introduction of image-text models [12, 19, 28]. The image-text models align images and their text descriptions into a common space, which facilitates model generalization to unseen categories at inference time. However, it has been observed that downstream image classification accuracy highly depends on the specific input prompts. This motivates researchers to either fine-tune or auto-learn prompts when adapting multi-modal models to downstream vision tasks. [29, 30] propose CoOp and CoCoOp to automatically learn the prompt word embeddings in the few-shot settings, and show significant improvements over the vanilla zero-shot image classification based-on prompting. These are learning based approaches, requiring supervised data from downstream tasks, while our proposed method is zero-shot and post-hoc without using any supervised data. In concurrent work, [24] proposes learning prompt embeddings in an unsupervised manner by minimizing the entropy of the averaged prediction probability distribution, where each prediction is based on a random augmentation applied to the input image. Our work differs from [24] in the sense that we do not learn an input-dependent prompt embedding. Instead we only selectively modify the prompts using knowledge hierarchy for images that have unreliable predictions, and our modified new prompt is natural language rather than a numerical embedding.

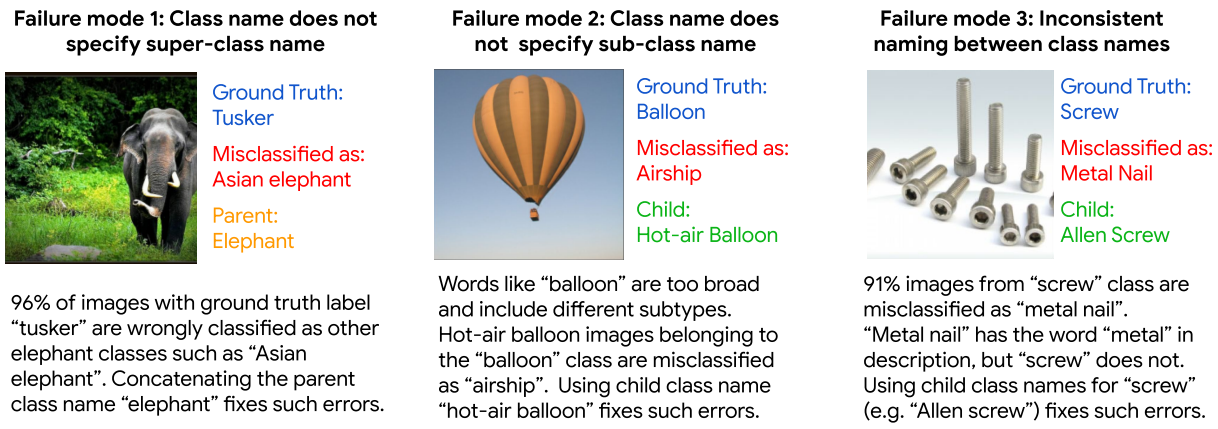

Figure 1. Typical failure modes in the cases where top-5 prediction was correct but top-1 was wrong.

Label hierarchy. Label hierarchy or label ontology are relational graphs among semantic labels. WordNet is one of the most widely used concept ontologies, and it has been used for visual recognition problems. Fergus et al. [5] leverage the WordNet hierarchy to define a semantic distance between any two categories and use this semantic distance to share labels. Deng et al. [3] propose a hierarchy and exclusion graph to explicitly model the semantic relations among labels, and significantly improve object classification by exploiting the rich label hierarchy. The idea of semantic distance defined on the WordNet ontology graph is also used in [22, 23] for transferring knowledge in zero-shot learning problems. We are similar to the above work in that we utilize the label semantics encoded by the label hierarchy as well, but label hierarchy in our case is used in the multi-modality scenarios: textual labels and visual images are represented in the same latent space, therefore, the hierarchy structure is directly exploited in the representation space to steer the recognition process.

3. Zero-shot inference failure case analysis

Given that the top-1 accuracy (64.2%) is much lower than top-5 accuracy (89.4%) for zero-shot ImageNet classification using CLIP, we investigated the failure cases that are "top-5 correct but top-1 wrong" (12605 images, 25.2% of all test images). Table. 1 in Suppl. shows some representative classes. The failure modes are summarized as:

(1) Class name does not specify super-class name: Some classes, whose class names do not have their WordNet ancestor (e.g., "tusker", one of 1k ImageNet classes, does not have its parent "elephant" in the class name), may have a relatively lower score than other classes, which explicitly have the ancestor present in the class name (e.g., "Asian elephant"). See examples in Fig. 1 (Left).

(2) Class name does not specify sub-class name: If the class name is too abstract, then its CLIP embedding is not necessarily close to the image embedding: e.g, CLIP wrongly classifies most images from "balloon" class as airship, see Fig. 1 (Middle). That is because there are distinct kinds of balloons, each belonging to a different semantic subgroup. Relying on the text embedding of the fine-grained children's class names (e.g., using "hot-air balloon") often fixes these errors. [1] reported the similar issue of label ambiguity in ImageNet.

(3) Inconsistent naming between class names: Some ImageNet class names are nouns, but others are adjective-prefixed nouns. This may make CLIP text embedding biased, see one example in Fig. 1 (Right) where images from "screw" class are misclassified as "metal nail".

4. Proposed Method

As shown in Section 3, CLIP models can be sensitive to different text prompts for images in certain classes. In this section, we first propose a confidence estimation method to identify low confidence predictions. We show that the identified subset has much lower accuracy than the average (Sec.4.1). We next develop a principled method that utilizes knowledge hierarchy to improve the accuracy of the low confidence subset, and consequently improve the overall accuracy on the whole datasets (Sec. 4.2).

4.1. Self-consistent zero-shot confidence estimation

Given an image $\boldsymbol{x}$ and a candidate class name c, where $c \in \mathcal{C}, |\mathcal{C}| = 1000$, the CLIP model encodes $\boldsymbol{x}$ and c respectively by its image encoder f_{image} and text encoder f_{text}, denoted as $\boldsymbol{z}_m = f_{image}(\boldsymbol{x})$ and $\boldsymbol{z}_c = f_{text}(c)$. The prediction logit score is defined as $\text{logit}(\boldsymbol{x}, c) = \cos(\boldsymbol{z}_m, \boldsymbol{z}_c)$, where $\cos(\cdot, \cdot)$ is the cosine similarity between two vectors, and the predicted class is $\arg\max_{c \in \mathcal{C}} \text{logit}(\boldsymbol{x}, c)$. We estimate the confidence by the self-consistency rate when applying different context prompts and image augmentations.

Confidence estimation via text prompts. To improve the zero-shot classifier's performance, the CLIP paper [19] hand crafted various context prompts, e.g. "`A photo of a big {label}`" and "`A photo of a`

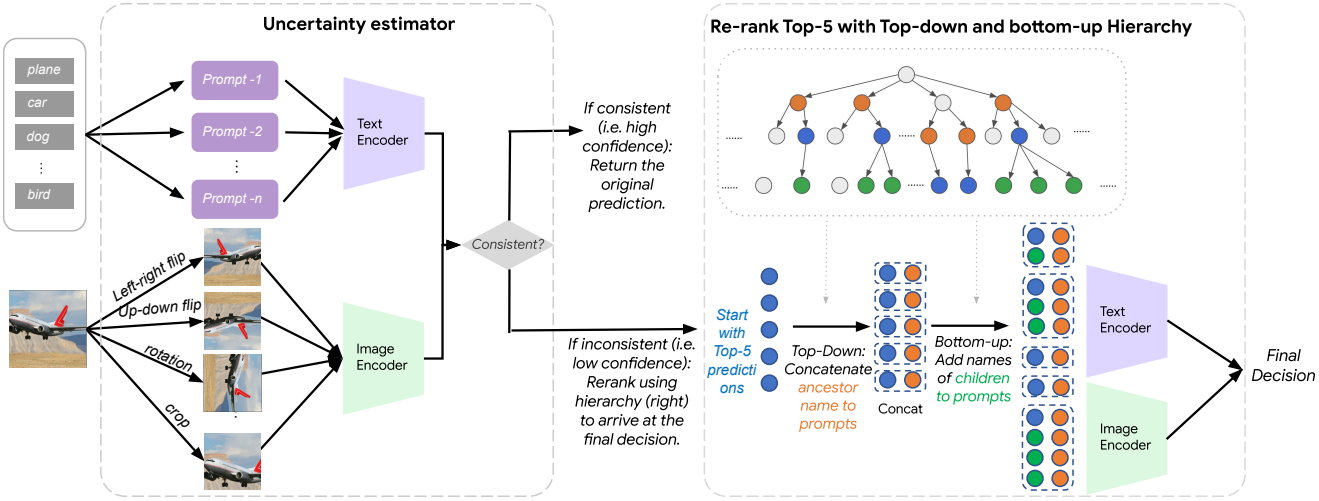

Figure 2. Our zero-shot classification pipeline consists of 2 steps: confidence estimation via self-consistency (left block) and top-down and bottom-up label augmentation using the WordNet hierarchy (right block). See Algorithms 1 and 2 for pseudocode.

`small {label}`"), for different datasets for the purpose of prompt ensembling: For an image $\boldsymbol{x}$, given a set of context prompts $\mathcal{T}$, the ensembled logit score is $\text{logit}(\boldsymbol{x}, \mathcal{T}(c)) = \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} \text{logit}(\boldsymbol{x}, t(c))$, where $t(c)$ denotes the new prompt after applying context prompt $t(\cdot)$ to c. Here instead of using the prompts for ensembling, we make use of the prompts to define our confidence score. Given a set of prompts $\mathcal{T}$, we apply each of the prompt $t(\cdot)$ for the classifier, and see if the top-1 prediction is the same as that when applying no prompt. We use the percentage of prompts that have consistent top-1 prediction with that without prompt as the confidence score $S_{\mathcal{T}}(\boldsymbol{x})$, i.e.

$$S_{\mathcal{T}}(\boldsymbol{x}) = \frac{\sum_{t\in\mathcal{T}} \mathbb{1}\{\hat{c}(\boldsymbol{x}, t) = \hat{c}(\boldsymbol{x}, \emptyset)\}}{|\mathcal{T}|} \tag{1}$$

where $\hat{c}(\boldsymbol{x}, \emptyset) = \arg\max_{c\in\mathcal{C}} \text{logit}(\boldsymbol{x}, c)$ is the top-1 prediction using the pure class name, and $\hat{c}(\boldsymbol{x}, t) = \arg\max_{c\in\mathcal{C}} \text{logit}(\boldsymbol{x}, t(c))$ is the top-1 prediction when applying prompt $t(\cdot)$. Intuitively, a reliable prediction should have highly consistent top-1 predictions when context prompts are applied or not, and therefore should have a high confidence score $S_{\mathcal{T}}(\boldsymbol{x})$ with respect to the prompt set $\mathcal{T}$, and vice versa.

Confidence estimation via image perturbation. We can also estimate the confidence of a prediction based on the self-consistency when applying different perturbations to the input image. Intuitively, if the top-1 predictions are inconsistent when applying different image perturbations, the prediction is unreliable. Specifically, we consider the common image transformations, left-right flip, rotation, crop, etc., and apply the perturbation method $b(\cdot)$ to the input image, and infer the predicted class as $\hat{c}(\boldsymbol{x}, b) = \arg\max_{c\in\mathcal{C}} \text{logit}(b(\boldsymbol{x}), c)$. We define the confidence score with respect to a set of image perturbations $\mathcal{B}$ as,

$$S_{\mathcal{B}}(\boldsymbol{x}) = \frac{\sum_{b\in\mathcal{B}} \mathbb{1}\{\hat{c}(\boldsymbol{x}, b) = \hat{c}(\boldsymbol{x}, \emptyset)\}}{|\mathcal{B}|} \tag{2}$$

We expect a high confidence prediction to have highly consistent prediction when applying different image perturbations, and therefore to have a high confidence score $S_{\mathcal{B}}(\boldsymbol{x})$ with respect to the image perturbation set B.

Determining the low confidence subset by combining the two confidence estimations. The confidence score we proposed in Eq. (1) and Eq. (2) are continuous values. A threshold needs to be determined if we want to select a subset of examples with low confidence using the continuous confidence score. In practice, the threshold can be chosen based on the requirement of recall and precision trade-off in the real application. In our study, to bypass the threshold selection, we propose to use a binary criterion for determining the low confidence set.

For IamgeNet dataset, the CLIP paper [19] designed total 80 context prompts. We define four sets based on the 80 prompts: the first 40 prompts $\mathcal{T}_1$, the last 40 prompts $\mathcal{T}_2$, all 80 prompts $\mathcal{T}_3$, and no prompts $\mathcal{T}_4 = \emptyset$. We apply the four different sets of prompts to the classifier and see if their top-1 predictions are all consistent or not, i.e. $\hat{c}(\boldsymbol{x}, \mathcal{T}_1) = \hat{c}(\boldsymbol{x}, \mathcal{T}_2) = \hat{c}(\boldsymbol{x}, \mathcal{T}_3) = \hat{c}(\boldsymbol{x}, \mathcal{T}_4)$. Then we determine the low confidence subset $\mathcal{O}_{\mathcal{T}}$ as those examples who have inconsistent predictions among the 4 prompts sets. We studied other choices such as using a random set of 40 prompts as $\mathcal{T}_1$, or splitting the 80 prompts into more subgroups, and found the results were very similar.

Similarly we also determine a low confidence subset $\mathcal{O}_{\mathcal{B}}$ based on image perturbations. In practice we found left-right flip works the best among the above mentioned perturbations. Thus for simplicity, we compare the top-1 prediction when applying the left-right flip to the input image and

Algorithm 1: Zero-shot confidence estimation

Input: Input images $\mathcal{X} = \{\boldsymbol{x}_i\}_{i=1}^N$, Candidate class set $\mathcal{C}$, image encoder f_{image} and text encoder f_{text}, text threshold τ_t, image threshold τ_i

Output: Low confidence set $\mathcal{O}$

Low confidence set $\mathcal{O}_\mathcal{T} \leftarrow \emptyset$ ▷Confidence estimation via text prompts

Sample L different context prompt $t_1, t_2 \dots t_L$

for $\boldsymbol{x}_i \in \mathcal{X}$ **do**

  Compute $S_\mathcal{T}(\boldsymbol{x}_i)$ based on Eq. (1)

  if $S_\mathcal{T}(\boldsymbol{x}_i) > \tau_t$ **then**

    $\boldsymbol{x}_i$ has high confidence prediction

  else

    $\mathcal{O}_\mathcal{T} \leftarrow \mathcal{O}_\mathcal{T} \cup \boldsymbol{x}_i$

Low confidence set $\mathcal{O}_\mathcal{B} \leftarrow \emptyset$ ▷Confidence estimation via image perturbation

Sample M perturbation methods $b_1, \dots, b_M$

for $\boldsymbol{x}_i \in \mathcal{X}$ **do**

  Compute $S_\mathcal{B}(\boldsymbol{x}_i)$ based on Eq. (2)

  if $S_\mathcal{B}(\boldsymbol{x}_i) > \tau_i$ **then**

    $\boldsymbol{x}_i$ has high confidence prediction

  else

    $\mathcal{O}_\mathcal{B} \leftarrow \mathcal{O}_\mathcal{B} \cup \boldsymbol{x}_i$

$\mathcal{O} \leftarrow \mathcal{O}_\mathcal{T} \cup \mathcal{O}_\mathcal{B}$

the top-1 prediction when using raw image. If their predictions are not consistent, that example will be included into the low confidence set $\mathcal{O}_\mathcal{B}$.

Finally, we use the union of the two low confidence sets $\mathcal{O}_\mathcal{T}$ identified using the text prompts and and $\mathcal{O}_\mathcal{B}$ identified using the image perturbations as the final low confidence subset $\mathcal{O}$ in the following experiments. Algorithm 1 shows the low confidence set generation process.

4.2. Top-down and bottom-up label augmentation using WordNet hierarchy

Through extensive analysis of the incorrect predictions among the identified unreliable predictions, we found that many of them are caused by CLIP's lack of robustness to prompts. Instead of tuning the prompt templates, we focus on how to augment `{label}` in "`A photo of a {label}`". A proper prompt that specifies both the generic type and the more specific sub-types of this class are very important for correctly classifying the image. However, the ImageNet [4] class names are not all defined with similar specificity and some classes are more abstract than others, e.g. 350 classes have children, while the rest of the classes have no children. See Suppl. Fig. 1 for more details. To make the ImageNet classification problem better suited to CLIP, we leverage the underlying WordNet hierarchy and develop a top-down and bottom-up class name augmentation method to improve zero-shot prediction accuracy for unreliable predictions.

The WordNet hierarchy is a semantic concept ontology, with nodes being cognitive synonyms indicating different concepts, and edges indicating the super-subordinate relation between concepts. Traveling upward from leaf nodes to the root, the concepts start from the very specific to the generic. For example, starting from the edge node "strawberry" to the root are "berry", "edible fruit", "produce", "food", "solid", "matter", and "physical entity" (the root). As we have seen in the failure mode analysis, many of the imageNet class names suffer from either being too abstract or being too specific, so that their concepts do not align well with the visual concepts the CLIP model learned in training. We propose using the WordNet knowledge hierarchy to augment the class labels in prompts so that the CLIP model has a better match between the image and prompts.

Top-down: augmenting class names with parent. As shown in failure case analysis, adding the super-class name to reduce ambiguity and to encourage the model's attention on the generic concept is helpful for improving the accuracy. Therefore we propose using WordNet to find the parent node of the raw class name, and concatenate it to the class name, i.e. $\text{logit}(\boldsymbol{x}, c) = \text{logit}(\boldsymbol{x}, [c; p(c)])$ where $p(c)$ is the parent node's name of the class name c, and $[c; p(c)]$ means the string concatenation of the class name and the parent name. We apply the method to top-5 predicted classes. Using the newly defined class names, we are able to re-rank the top-5 predictions for the identified unreliable subset of images. Note that WordNet contains a few very abstract class names for nodes, such as "physical entity", "artifact", "matter", etc. We found that such parent nodes are not informative, hence we remove them. There are also many academic words in WordNet, for example the parent node of sea anemone is "anthozoan", which can be rare in CLIP training data. Adding those academic words to class name makes the prediction even less robust. So we simplify the WordNet by pruning based on an estimation of the word frequency in CLIP training data by using embedding norm.

Bottom-up: augmenting class names with children. Some ImageNet class names are generally abstract, but the ImageNet images may belong to a specific subtype of the class. For example, "balloon" is a class name in ImageNet, but most balloon images in ImageNet are actually "hot-air balloon", which is a child of "balloon" in WordNet hierarchy. The logit score for a parent class is not necessarily higher than the score for its child classes, mismatching with hierarchy prior. To accurately classify the images using CLIP, we need to augment the class name with fine-grained child subclasses. For each class c having children in the WordNet hierarchy, we redefine the logit score as the max score over itself and all its children, i.e., $\text{logit}(\boldsymbol{x}, c) = \max\{\text{logit}(\boldsymbol{x}, c), \text{logit}(\boldsymbol{x}, c_1), \dots, \text{logit}(\boldsymbol{x}, c_r)\}$, where $c_1 \dots c_r$ are the r children of the node c in the WordNet

Algorithm 2: Top-down and bottom-up class label augmentation using WordNet hierarchy

Input: Input image $\boldsymbol{x} \in \mathcal{O}$, top-5 candidate class set $\mathcal{C}_{top5}$, sparse WordNet hierarchy H, image encoder f_{image} and text encoder f_{text}

Output: Predicted class of $\boldsymbol{x}$

Candidate class set $\mathcal{C} \leftarrow \emptyset$

for $c \in \mathcal{C}_{top5}$ **do**

  $\mathcal{C} \leftarrow \mathcal{C} \cup [c; \text{parent}(c)]$, where parent($c$) is the parent of c in H ▷Top-down

  if c has $r \geq 1$ children $c_1 \dots c_r$ *in* H **then**

    $\mathcal{C} \leftarrow \mathcal{C} \cup \{[c_j; \text{parent}(c)]\}_{j=1}^{r}$ ▷Bottom-up

$\hat{c} \leftarrow \arg\max_{c \in \mathcal{C}} \text{logit}(\boldsymbol{x}, c)$

if $\hat{c} \in \mathcal{C}_{top5}$ **then**

  final prediction $\leftarrow \hat{c}$

else

  final prediction $\leftarrow$ parent($\hat{c}$)

hierarchy. We apply this bottom-up method to top-5 predicted class names, and re-rank the top predictions.

Combining Top-down and bottom-up. In practice, we use both children and the ancestor(parent) to augment each class c, to transfer semantic information bidirectionally in both top-down and bottom-up way: the ancestor(parent) class is more generic than c, and has better chance to disambiguate instance from a more abstract level; on the other hand, children categories have more specific attribute description, and the attribute descriptions are semantically meaningful representations bridging the gap between the image embedding and its abstract class concept c. Then the final logit score between x and c is:

$$\begin{aligned}\text{logit}(\boldsymbol{x}, c) = \max\{&\text{logit}(\boldsymbol{x}, [c; p(c)]), \\ &\text{logit}(\boldsymbol{x}, [c_1; p(c)]), \dots, \text{logit}(\boldsymbol{x}, [c_r; p(c)])\}\end{aligned} \quad (3)$$

where $p(c)$ is parent of c, and $c_1 \dots c_r$ are c's children. The $\hat{c}$, where $\hat{c} \in \mathcal{C}_{top5}$, with the maximal logit score is the predicted class of $\boldsymbol{x}$. See Algorithm 2 for details.

5. Experiments and Results

Our proposed method is composed of two steps and we conduct experiments to verify the effectiveness of each step: (1) Use zero-shot confidence estimation to identify the low confidence subset of samples (see Fig. 3 for the results), and (2) Augment the class label using top-down and bottom-up strategies based on the sparsified WordNet on the low confidence subset to improve the accuracy (See Table 1 and Table 2 for the results).

5.1. Our proposed confidence score is better suited for selective prediction than baselines

A well-calibrated confidence estimator should score high for those correct predictions, and low for incorrect predictions. As a result, a good confidence estimator should be a good predictor for prediction correctness. We plot the receiver operating characteristic (ROC) curve and compute the area under the ROC curve (AUC) as a quantitative measure to compare our proposed confidence estimation with the baselines. An AUROC of 1.0 indicates perfect separation between correct and incorrect predictions, and 0.5 means the two groups are not distinguishable. Maximum logit score, $\max_{c \in \mathcal{C}} \text{logit}(\boldsymbol{x}, c)$ is one of the most commonly used confidence score for classification problems in single modal models [8], so we consider it as our baseline. Fig. 3a and 3c clearly show that our confidence score is significantly better than the baseline method at distinguishing between correct and incorrect predictions, for both CLIP and LiT models. The AUC score for our proposed method is above 0.8 while that for the baseline method is around 0.7.

We also compare our method with the baseline in the scenario of selective prediction. Given a budget of abstention rate $\alpha\%$, the best strategy is to abstain the $\alpha\%$ samples with the lowest confidence scores. If the confidence score is well calibrated, the accuracy for the abstained set will be low and as an evidence the accuracy of the remaining set would be high. We plot the selective prediction curves [14], which reports the accuracy on the remaining set as a function of the abstention rate. Fig. 3b and 3d show that our proposed confidence score results in higher accuracy than the baseline maximum logit score at all abstention rates for both CLIP and LiT.

Prompt ensemble has been shown to improve accuracy and robustness of the prediction, so here we also compare ours with the maximum logit score after applying prompt ensemble. As shown in the selective prediction curves, although the prompt ensemble indeed helps to achieve higher accuracy (dashed line) than that using the pure class name (solid line), it is still inferior to our proposed method.

5.2. Using hierarchy to help improve zero-shot accuracy on low confidence subset

Using top-down and bottom-up label augmentation significantly improves the accuracy on the low confidence subset. We apply the top-down and bottom-up label augmentation on the low confidence subset: to better combine child and parent name, we create a prompt template to transform the child and parent name pairs into a new class name $\tilde{c}$ in natural language: "{child} which is a kind of {parent}" (different prompt templates may have different results). Table 1 shows improvement of 17.13% on the top-1 accuracy (from 21.58% to 38.71%) for the identified low confidence subset of samples, and overall 3.6% on the top-1 accuracy (64.18% to 67.78%) for all samples in ImageNet. We show similar improvement on the zero-shot accuracy for ImageNet shifted datasets. To investigate if our method works for other multi-modal models, we apply it to the LiT [28] model and observe that our method improves accuracy

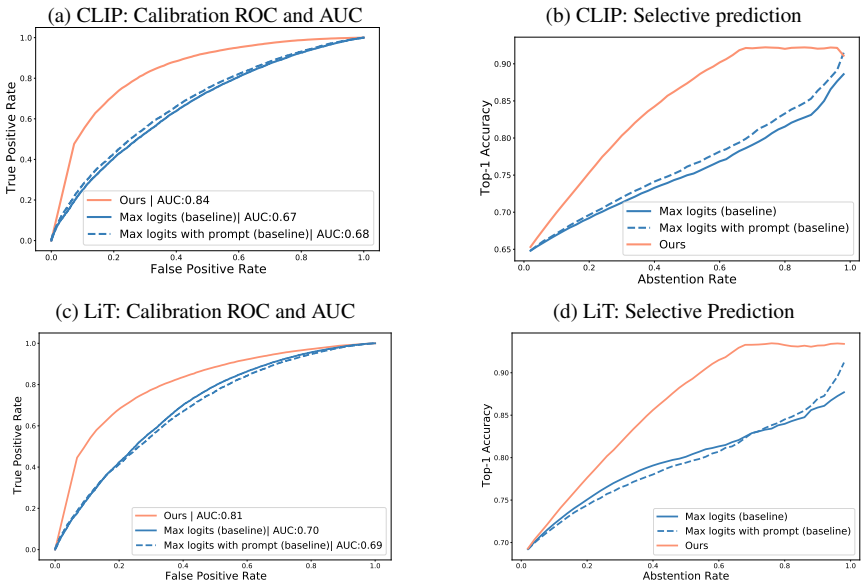

Figure 3. ROC plots (left column) show that our proposed confidence score is better at distinguishing correct and incorrect predictions and results in higher AUC scores than baselines for both CLIP (ViT-B/16) (a) and LiT (ViT-B/32)(c). Selective prediction curves (right column) show that our proposed confidence score is better at abstaining incorrect predictions and as a result the accuracy of the remaining set is higher than the baselines for both CLIP (ViT-B/16) (b) and LiT (ViT-B/32) (d).

Table 1. CLIP (ViT-B/16) and LiT (ViT-B/32) zero-shot top-1 accuracy comparison between baseline and ours (w/ hierarchy).

		CLIP	(Ours) Hierarchy-CLIP	LiT	(Ours) Hierarchy-LiT
ImageNet [4]	Low conf. set	21.58%	**38.71%**	31.18%	**37.25%**
	Full set	64.18%	**67.78%**	68.26%	**69.41%**
ImageNet-v2 [21]	Low conf. set	17.77%	**32.50%**	27.08%	**31.45%**
	Full set	58.06%	**61.07%**	60.11%	**61.11%**
ImageNet-R [9]	Low conf. set	16.79%	**27.91%**	21.82%	**22.93%**
	Full set	56.88%	**59.46%**	66.54%	**66.75%**
ImageNet-Adversarial [11]	Low conf. set	10.13%	**18.44%**	7.19%	**8.95%**
	Full set	26.12%	**29.23%**	13.93%	**14.56%**
ImageNet-Sketch [25]	Low conf set	13.74%	**23.18%**	21.51%	**24.42%**
	Full set	44.71%	**47.28%**	52.47%	**53.17%**

for LiT models as well. See Supp. Fig. 2 for qualitative visualization.

Generalizability to non-ImageNet datasets To show the generalizability of our methods on non-ImageNet datasets, We conducted experiments on 4 additional datasets: Caltech-101 [15] (101 categories), Flower-102 [17] (102 flower categories), Food-101 [2] (101 food categories) and Cifar-100 [13] (100 categories). For each dataset, a subset of their categories are exist/aligned with WordNet hierarchy, we only apply our method on those WordNet aligned class names, where we could find their ancestor and children. We keep the other class names unmodified. We use CLIP (ViT-B/16) as multi-modal model. Table 2 shows that our method consistently improved accuracy on the low-confidence set (low) and the entire set (full):

Table 2. Generalizability to non-ImageNet datasets (CLIP (ViT-B/16) zero-shot top-1 accuracy).

Dataset	orig (low)	ours (low)	orig (full)	ours (full)
Caltech-101 [15]	10.6 %	27.2% (+16.6%)	74.1%	77.1% (+3.0%)
Flower102 [17]	20.0%	29.4% (+9.4%)	63.7%	65.3% (+1.6%)
Food-101 [2]	28.2%	49.0% (+20.8%)	84.7%	86.8% (+2.1%)
Cifar-100 [13]	9.4%	17.5% (+8.1%)	31.8%	35.2% (+3.4%)

5.3. Ablation study

Generalizability to other backbones To study the generalization of our method to different model architectures and sizes, we used 4 additional backbones of CLIP, including convolutional neural network (CNN) based backbones (ResNet-50, ResNet-101) and vision transformer (ViT) based backbones (ViT-B/32, ViT-B/16 and ViT-l/14). Table 3 shows the improved accuracy after using our method

Table 3. Generalizability to different backbones with CLIP.

backbone	ResNet-50	ResNet-101	ViT-B/32	ViT-B/16	ViT-l/14
ACC (low)	+14.25%	+12.97%	+15.12%	+ 17.13%	+18.89%
ACC (full)	+3.73%	+3.71%	+3.65%	+ 3.60%	+3.23%

Table 4. CLIP (ViT-B-16) zero-shot top-1 accuracy comparison with prompt ensemble.

		Ensemble only	Hierarchy and Ensemble
ImageNet [21]	Low conf. set	41.05%	**42.09%**
	Full set	68.48%	**68.86%**
ImageNet-v2 [21]	Low conf. set	**36.39%**	36.34%
	Full set	**62.02%**	62.00%
ImageNet-R [9]	Low conf. set	35.13%	**36.12%**
	Full set	60.21%	**60.62%**
ImageNet-Adversarial [11]	Low conf. set	21.13%	**22.00%**
	Full set	30.59%	**31.07%**
ImageNet-Sketch [25]	Low conf. set	**27.13%**	26.56%
	Full set	**48.52%**	48.26%

on ImageNet with CLIP models of different backbones. Our method achieves consistently improved accuracies.

Our hierarchy-based label augmentation is complimentary to prompt ensembling. Prompt ensembling (PE) [19] requires a set of manually crafted prompt templates, and the zero-shot performance is sensitive to the set of prompts the model uses. Alternatively, our proposed method does not require a dedicated tuning of the prompt templates. We directly augment the class name with knowledge of the hierarchy from WordNet. In addition, PE is computationally intensive because it needs to infer the embeddings of 80 prompt templates where each is applied with 1000 ImageNet classes, while our method only need to infer once for each of the predicted top-5 labels. Our method is more straightforward and interpretable given that it clearly shows the contribution of parent/child in the decision. Intuitively, PE is typically focused on fixing $\{class\}$ and augmenting contextual templates, while our method augments the $\{class\}$ with a fixed contextual template. To verify if our hierarchy-based method is complimentary with prompt ensembling, we apply prompt ensembling after applying our top-down and bottom-up label augmentation. For the low confidence set, we first create a prompt template to transform the child and parent name pairs into a new class name $\tilde{c}$ in natural language: "`{child} which is a kind of {parent}`". Then we apply the 80 prompts designed by the CLIP paper [19] individually to the new class name $\tilde{c}$, and then ensemble them. For the high confidence set, since we do not modify the class name using hierarchy information, we only apply the prompt ensemble. The performance is shown in Table 4. We compare the zero-shot accuracy using the vanilla prompt ensembling method proposed in CLIP, and the zero-shot accuracy using our combined version of hierarchy-based class name augmentation and prompt ensembling. As shown in the table, using both hierarchy and prompt ensembling achieves better or on par accuracy with the prompt ensemble alone, suggesting that the two methods can be combined. Considering the prompt ensemble requires manually designed prompt templates and much greater inference time, our hierarchy-based class name augmentation is simple, efficient and effective. We also computed IoU of corrected low-confidence instances (*low set*) between PE and our method: the IoU is 0.55, which implies the two methods are complementary for fixing errors.

Table 5. Effect of threshold of confidence score on zero-shot accuracy.

Threshold	Low conf. set size	Acc on low conf. set	Acc on full set
0.47	10000	19.40%	68.72%
0.52	11000	20.82%	68.78%
0.57	12000	22.06%	68.82%
0.62	13000	23.58%	68.85%
0.66	14000	25.01%	**68.88%**
0.70	15000	26.51%	68.86%

Effect of threshold of confidence score on zero-shot accuracy. In Table 1 we use a binary criterion to determine the low confidence set. We can alternatively use the continuous confidence score by choosing a threshold based on the trade-off between precision and recall. Changing the threshold of the confidence score can lead to different numbers of samples in the low confidence set. We study the effect of threshold on zero-shot accuracy. Table 5 shows the overall accuracy with different thresholds. We find that the overall accuracy is relatively robust to the threshold selection, in the wide range from 0.47 to 0.70.

6. Conclusion

Multi-modal models' generalization and robustness is critical for deployment. Motivated by the big gap between top-1 and top-5 accuracy in ImageNet zero-shot classification, we investigated the failure modes and found that the model's prediction is very sensitive to text prompts. We describe a simple but efficient zero-shot post-hoc method to identify a subset of samples that are most likely to be predicted wrongly by a measure of self-consistency. For those in the low confidence subset, we use the WordNet hierarchy to augment class labels to enhance the robustness, resulting in up to 17.13% accuracy improvement on ImageNet. We show our method provides consistent improvement over other distribution shifted datasets (ImageNet variants), four other datasets, and is generalizable to other image-text models and different backbones.

Acknowledgments This work was supported in part by C-BRIC (one of six centers in JUMP, a Semiconductor Research Corporation (SRC) program sponsored by DARPA), DARPA (HR00112190134) and the Army Research Office (W911NF2020053). The authors affirm that the views expressed herein are solely their own, and do not represent the views of the United States government or any agency thereof.

References

[1] Lucas Beyer, Olivier J Hénaff, Alexander Kolesnikov, Xiaohua Zhai, and Aäron van den Oord. Are we done with imagenet? *arXiv preprint arXiv:2006.07159*, 2020. 1, 3

[2] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014. 7

[3] Jia Deng, Nan Ding, Yangqing Jia, Andrea Frome, Kevin Murphy, Samy Bengio, Yuan Li, Hartmut Neven, and Hartwig Adam. Large-scale object classification using label relation graphs. In *ECCV*, pages 48–64, 2014. 3

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1, 5, 7

[5] Rob Fergus, Hector Bernal, Yair Weiss, and Antonio Torralba. Semantic label sharing for learning with many categories. In *ECCV*, pages 762–775, 2010. 3

[6] Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. arxiv e-prints, page. *arXiv preprint arXiv:1506.02142*, 3, 2015. 2

[7] Yunhao Ge, Jiashu Xu, Brian Nlong Zhao, Laurent Itti, and Vibhav Vineet. Dall-e for detection: Language-driven context image synthesis for object detection. *arXiv preprint arXiv:2206.09592*, 2022. 1

[8] Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. *arXiv preprint arXiv:1911.11132*, 2019. 2, 6

[9] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *ICCV*, 2021. 1, 7, 8

[10] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016. 2

[11] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, 2021. 1, 7, 8

[12] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, pages 4904–4916, 2021. 1, 2

[13] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 7

[14] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *NeurIPS*, 30, 2017. 2, 6

[15] FF Li, M Andreetto, MA Ranzato, and P Perona. Caltech101. *Computational Vision Group, California Institute of Technology*, 2003. 7

[16] Jeremiah Zhe Liu, Shreyas Padhy, Jie Ren, Zi Lin, Yeming Wen, Ghassen Jerfel, Zack Nado, Jasper Snoek, Dustin Tran, and Balaji Lakshminarayanan. A simple approach to improve single-model deep uncertainty via distance-awareness. *arXiv preprint arXiv:2205.00403*, 2022. 2

[17] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008. 7

[18] Hieu Pham, Zihang Dai, Golnaz Ghiasi, Kenji Kawaguchi, Hanxiao Liu, Adams Wei Yu, Jiahui Yu, Yi-Ting Chen, Minh-Thang Luong, Yonghui Wu, et al. Combined scaling for open-vocabulary image classification. *arXiv preprint arXiv:2111.10050*, 2021. 1

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 1, 2, 3, 4, 8

[20] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 1

[21] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *ICML*, 2019. 1, 7, 8

[22] Marcus Rohrbach, Michael Stark, and Bernt Schiele. Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In *CVPR*, pages 1641–1648, 2011. 3

[23] Marcus Rohrbach, Michael Stark, György Szarvas, Iryna Gurevych, and Bernt Schiele. What helps where–and why? semantic relatedness for knowledge transfer. In *CVPR*, pages 910–917, 2010. 3

[24] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *arXiv preprint arXiv:2209.07511*, 2022. 2

[25] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *NeurIPS*, pages 10506–10518, 2019. 1, 7, 8

[26] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022. 2

[27] Yeming Wen, Dustin Tran, and Jimmy Ba. Batchensemble: an alternative approach to efficient ensemble and lifelong learning. *arXiv preprint arXiv:2002.06715*, 2020. 2

[28] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *CVPR*, pages 18123–18133, 2022. 1, 2, 6

[29] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, pages 16816–16825, 2022. 2

[30] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *IJCV*, 130(9):2337–2348, 2022. 2