Supplementary Material for

Benchmarking Zero-Shot Recognition with Vision-Language Models: Challenges on Granularity and Specificity

Zhenlin Xu[1]* Yi Zhu[2]† Siqi Deng[1] Abhay Mittal[3]† Yanbei Chen[1] Manchen Wang[3]
Paolo Favaro[1] Joseph Tighe[3]† Davide Modolo[1]
[1]AWS AI Labs [2]Boson AI [3]Meta

A. A Two-level Granularity Benchmark

In this section, we presents an simplified granularity benchmark with two-levels of semantic hierarchy. The results are consistent with our observations in the main paper.

Two-level Dataset Our evaluation starts on a dataset with two levels of labels: N_{cg} coarse-grained (CG) classes $Y_{cg} = \{y^i_{cg}\}$, where $i \in \{1, ..., N_{cg}\}$, and each CG class has N^i_{fg} fine-grained (FG) children classes $Y^i_{fg} = \{y^{i,j}_{fg}\}$, where $j \in \{1, ..., N^i_{fg}\}$. In total, there are $N_{fg} = \sum_{i=1}^{N_{cg}} N^i_{fg}$ FG classes. To create our two-level classification dataset, we adapt the tiered-ImageNet [1] benchmark, which has 608 FG classes (a subset of the original 1000 classes of ImageNet-1K) organized under 34 CG classes and covers 30,400 out of 50,000 ILSVRC-12 validation images.

Evlauation protocol For *two-level* granularity, we measure the performance difference of CG classification between using direct predictions with CG prompts and propagated FG predictions. The simplest propagation method is to assign the predicted FG labels to their CG parents' labels. For instance, if an image is predicted as ”golden retriever” in the FG classification, it is labeled with its CG parent class ”animal.” Intuitively, if a model exhibits consistent understanding of CG and FG concepts, the performance of CG classification using CG prompts should be similar to propagating the results from FG classification. An alternative way of propagating FG to CG concepts is using the aggregated embeddings of FG prompts for CG classifcation. Specifically, for the i-th CG class, we compute the average of the FG prompt embeddings as the CG prompt embedding: $E^{prop}_t(y^i_{cg}) = \frac{1}{N^i_{fg}} \sum_{j=1}^{N^i_{fg}} E_t(y^{i,j}_{fg})$. We use top1 accuracy as the classification metric.

B. A Language Only Study

In the main paper, we have highlighted the issues faced by vision and language models (VLMs) in zero-shot recognition tasks, focusing on both granularity and correctness analyses. Since these analyses primarily involve working with different text inputs while keeping the visual inputs constant, improving the language encoder becomes a natural next step. We address the question of whether language embeddings from pre-trained large-scale language models (LLMs) exhibit better behavior compared to VLMs. To investigate this, we design a language-only task.

Specifically, we conduct a text classification task that involves classifying fine-grained (FG) concepts to their corresponding coarse-grained (CG) concepts using the same two-level ImageNet dataset as in Section 4.1. This results in a 34-way classification task with 608 text samples (FG concept prompts). Similar to zero-shot image classification, we compute the cosine similarity between the language embeddings of FG and CG prompts and classify a FG concept to the CG concept with the highest similarity score. To incorporate the generative model GPT-3 for this task, we design the following zero-shot prompt:

> ”Classify a given concept into one of the following classes: ${*all coarse-grained concepts* }.
> Q: ${*a fine-grained_concept*} A:”

Tab. 2 Presents the performance of LLMs[1] or the language encoder of VLMs on the language-only task. Surprisingly, LLMs, even when fine-tuned for sentence embedding, do not outperform the language encoder of VLMs. However, we find that GPT-3 performs significantly better in a generative manner. This suggests that when dealing with concept relationships on a larger scale where simple embedding similarity struggles, generative modeling may offer a more powerful approach to capture complex semantic knowledge and model the relationships effectively.

*Correspondence to: xzhenlin@amazon.com

†Work done while at Amazon

[1] We use pretrained models provided by sentence-transformer

Table 1. Evaluating vision-Language model zero-shot classification performance (top-1 accuracy) on fine-grained classes (FG) and coarse-grained (CG) classes. The CG classification results are obtained through two methods: relating predicted FG class labels to their CG parents ($\text{CG}_{\text{FG-label}}$) and using the average of the FG prompt embeddings as the CG prompt embedding ($\text{CG}_{\text{FG-emb}}$). We measure the differences (Δ) with CG classification using CG class prompts ($\text{CG}_{\text{direct}}$), which reveals the discrepancy in CG-FG performance of vision-language models.

Model	Arch	Training data	$\text{FG}_{\text{direct}}$	$\text{CG}_{\text{direct}}$	$\text{CG}_{\text{FG-label}}$ (Δ)	$\text{CG}_{\text{FG-emb}}$ (Δ)
CLIP	ViT-B-32	Private400M	66.47	50.15	86.35 (+36.2)	72.62 (+22.47)
Open-CLIP	ViT-B-32	LAION400M	63.82	35.98	84.08 (+48.1)	69.65 (+33.67)
	ViT-B-32	LAION2B	69.78	45.54	87.39 (+41.85)	71.54 (+26)
	ViT-L-14	LAION2B	77.72	49.74	91.83 (+42.09)	76.49 (+26.75)
	VIT-H-14	LAION2B	80.39	52.22	92.86 (+40.64)	77.43 (+25.21)
UniCL	Swin-B	YFCC14M	41.14	37.37	69.67 (+32.3)	59.75 (+22.38)
	Swin-B	IN21K	30.6	53.14	66.26 (+13.12)	59.5 (+6.36)
	Swin-B	IN21K+YFCC14M	45.91	52.27	76.84 (+24.57)	67.63 (+15.36)
	Swin-B	IN21K+YFCC14M+GCC15M	60.17	51.9	83.44 (+31.54)	68.37 (+16.47)
K-LITE	Swin-B	IN21K+YFCC14M+GCC15M	54.75	44.92	81.85 (+36.93)	71.05 (+26.13)
BLIP	ViT-B-16	COCO+VG+CC+SBU +LAION+CapFilt-L	55.41	42.09	80.92 (+38.83)	69.69 (+27.6)
$\text{BLIP}_{\text{ft-coco}}$			58.02	46.75	84.7 (+37.95)	72.93 (+26.18)
FLAVA	ViT-B/16	PMD70M	59.48	50.11	83.37(+33.26)	70.84 (+20.73)

Table 2. Performance (accuracy) of classify a fine-grained concept to coarse-grain concept using language embedding models or generative language models.

Model Type	FG-to-CG Text Classification Accuracy (%)
CLIP-B	61.18
$\text{OpenCLIP-L}_{\text{LAION2B}}$	55.76
$\text{OpenCLIP-H}_{\text{LAION2B}}$	62.66
UniCL	52.96
KLITE	43.59
BLIP	50.00
FLAVA	57.40
all-roberta-large-v1	51.81
sentence-T5-large	52.47
sentence-T5-xl	55.26
$\text{GPT-3}_{\text{text-davinci-002}}$	71.17

C. Limitations of Our Study

While our study provides valuable insights into the challenges and limitations of vision-and-language models (VLMs) for zero-shot visual recognition, it is important to acknowledge several limitations. Firstly, our experiments primarily focus on a specific set of VLMs, datasets, and evaluation metrics. While we have made efforts to select representative models and datasets, our findings may not fully generalize to the entire landscape of vision and language models. Generalizing the results to other VLM architectures or datasets requires further investigation and experimentation.

Secondly, our study is conducted within the context of the evaluation protocols and benchmarks we have proposed. While we have designed these protocols to address the challenges of zero-shot recognition in open-world settings, it is important to recognize that these benchmarks may not fully capture the complexities and variations present in real-world scenarios. Real-world applications may involve different

https://github.com/UKPLab/sentence-transformers types of data, varied distributions, and additional challenges that are not fully accounted for in our study.

Furthermore, the scalability of hard sample generation, as used in our fine-tuning experiments, presents a practical limitation. Generating diverse and representative hard positive and negative samples can be computationally expensive and time-consuming. Scaling up the generation process to cover a wide range of positive and negative cases with diverse variations poses a significant challenge and may require more efficient and scalable methods.

References

[1] Mengye Ren, Eleni Triantafillou, Sachin Ravi, Jake Snell, Kevin Swersky, Joshua B Tenenbaum, Hugo Larochelle, and Richard S Zemel. Meta-learning for semi-supervised few-shot classification. *arXiv preprint arXiv:1803.00676*, 2018. 1