Distilling Out-of-Distribution Knowledge from Large Language Models for CLIP Generalization

Supplementary Material

A. Dataset Details

The details of the 14 datasets used in our experiments are provided in Table 1. These include general object recognition datasets such as ImageNet [3] and Caltech101 [4]; fine-grained classification datasets like OxfordPets [11], StanfordCars [8], Flowers102 [10], Food101 [1], and FGVCAircraft [9]; a scene recognition dataset, SUN397 [15]; an action recognition dataset, UCF101 [13]; a texture classification dataset, DTD [2]; and a satellite imagery dataset, EuroSAT [5]. Additionally, ImageNet-A [6], ImageNet-R [7], ImageNet-V2 [12], and ImageNet-Sketch [14] are four variants of ImageNet.

Dataset	Classes	Train	Val	Test
ImageNet	1,000	1.28 M	N/A	50,000
Caltech101	100	4,128	1,649	2,465
OxfordPets	37	2,944	736	3,669
StanfordCars	196	6,509	1,635	8,041
Flowers102	102	4,093	1,633	2,463
Food101	101	50,500	20,200	30,300
FGVCAircraft	100	3,334	3,333	3,333
SUN397	397	15,880	3,970	19,850
DTD	47	2,820	1,128	1,692
EuroSAT	10	13,500	5,400	8,100
UCF101	101	7,639	1,898	3,783
ImageNetV2	1,000	N/A	N/A	10,000
ImageNetSketch	1,000	N/A	N/A	50,889
ImageNet-A	1,000	N/A	N/A	7,500
ImageNet-R	1,000	N/A	N/A	30,000

Table 1. Summary of the 14 datasets.

B. Effect of Description Numbers per Class

We investigate the influence of the number of detailed descriptions per category on model performance in the base-to-novel generalization setting, as presented in Table 2. By varying the number of descriptions, we examine the accuracy of both base and novel classes, along with the harmonic mean (HM). Overall, an increase in the number of descriptions leads to a gradual improvement in the HM. This can be attributed to the fact that a richer set of class-specific textual prior knowledge provides more valuable information to the model.

number	10	20	30	40	50	60	70
Base	83.01	82.79	81.15	81.06	**81.11**	81.35	81.41
Novel	64.56	65.27	67.94	70.15	**73.63**	72.18	72.24
HM	72.63	72.99	73.95	75.21	**77.18**	76.75	76.55

Table 2. Effect of the number of LLM-generated descriptions per category.

B.1. Robust of LLMs

To evaluate the robustness of our methods with respect to different LLMs, we also generate descriptions using GPT-4o, and the results are comparable. As shown in Table 3, using GPT-4o does not yield superior classification performance. This result may be attributed to two main reasons: 1. For common visual categories, GPT-3.5 is already capable of producing attribute-rich descriptions that are sufficiently aligned with CLIP's text encoder, resulting in diminishing returns when switching to a stronger language model. 2. CLIP's text encoder, trained on relatively simple natural language prompts, lacks deep semantic comprehension and is therefore limited in its ability to exploit the more expressive or nuanced content generated by GPT-4o.

Model	Base	Novel	HM
GPT-3.5	**81.11**	**73.63**	**77.18**
GPT-4o	81.62	73.09	77.11

Table 3. Performance based on different LLMs.

B.2. Samples of Descriptions

We select the first category from each dataset and provide one example description sentence per class. As shown in Table 4, most descriptions mention general visual concepts

such as color, size and texture, providing valuable cues for visual recognition. Additionally, some undesirable outputs are generated by the LLM, such as "I'm sorry, but I can't provide a description of images.", which can be considered as noise or redundancy within the textual priors. Our method does not perform additional filtering of these noisy or redundant descriptions. Instead, we leverage the attention mechanism to extract useful information from the entire set of descriptions, focusing on those that contribute to improving the model's generalization ability.

Dataset	Descriptions
ImageNet	"A tench looks like a freshwater fish with a dark olive-green back, fading to yellowish-brown on the sides ..."
Caltech101	"A face is the front part of a person's head, featuring the eyes, nose, mouth, cheeks, and chin ..."
OxfordPets	"The Abyssinian cat is an elegant and slender breed with a medium-sized, muscular body that is agile and graceful ..."
StanfordCars	"The 2000 AM General Hummer SUV is a robust and large vehicle known for its military-grade design..."
Flowers102	"A pink primrose has delicate, cup-shaped flowers with five softly rounded petals..."
Food101	"An apple pie is a classic dessert that consists of a flaky pastry crust filled with a sweet and aromatic mixture of ... "
FGVCAircraft	"The Boeing 707-320 is a four-engine commercial jet airliner that features a sleek and streamlined design..."
SUN397	"An abbey scene typically features a large, imposing stone building with intricate architecture..."
DTD	"A banded texture typically appears as distinct bands or stripes of different colors, patterns, or materials running parallel to each other on a surface..."
EuroSAT	"An Annual Crop Land in a centered satellite photo typically appears as a patchwork of vibrant green hues..."
UCF101	"When a person applies eye makeup, they typically start by holding a small hand mirror in one hand"

Table 4. Sample classes from each datasets and their corresponding descriptions provided by GPT-3.5-turbo.

References

[1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part VI 13*, pages 446–461. Springer, 2014. 1

[2] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014. 1

[3] 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

[4] Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, pages 178–178. IEEE, 2004. 1

[5] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. 1

[6] 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 *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8340–8349, 2021. 1

[7] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15262–15271, 2021. 1

[8] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013. 1

[9] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 1

[10] 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. 1

[11] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3498–3505. IEEE, 2012. 1

[12] 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

[13] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. A dataset of 101 human action classes from videos in the

wild. *Center for Research in Computer Vision*, 2(11):1–7, 2012. 1

[14] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. *Advances in neural information processing systems*, 32, 2019. 1

[15] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *2010 IEEE computer society conference on computer vision and pattern recognition*, pages 3485–3492. IEEE, 2010. 1