Task-Aware Clustering for Prompting Vision-Language Models

Supplementary Material

This supplementary material provides additional information about experiment setup, ablation studies, discussion, and source code. We respectfully note that the source code will be made publicly available.

A. Experiment setup

The evaluation is conducted in the base-to-novel, domain generalization, and cross-dataset transfer settings [7, 8, 18]. The average accuracy of three runs is reported as the final performance. The detailed descriptions of these settings are as follows.

Base-to-novel setting. In this setting, the classes of a dataset are divided into two disjoint sets, i.e., base and novel, and a model is optimized on the base set and evaluated on the base and novel sets. The involved standard benchmark datasets are: ImageNet [3], Caltech101 [10], OxfordPets [12], StanfordCars [9], Flowers102 [11], Food101 [1], FGVCAircraft [15], SUN397 [17], DTD [2], EuroSAT [4], and UCF101 [14]. The goal of this setting is to evaluate the resulting model's generalizability.

Domain generalization setting. In this setting, a model is optimized on ImageNet [3] and evaluated on ImageNet-A [6], ImageNet-R [5], ImageNet-Sketch [16], and ImageNetV2 [13]. The goal of this setting is to evaluate the resulting model's generalizability on datasets with domain shifts.

Cross-dataset transfer setting. In this setting, a model is optimized on ImageNet [3] and evaluated on Caltech101 [10], OxfordPets [12], StanfordCars [9], Flowers102 [11], Food101 [1], FGVCAircraft [15], SUN397 [17], DTD [2], EuroSAT [4], and UCF101 [14]. The goal of this setting is to evaluate the resulting model's transferability.

Please refer to [18] for the instructions to prepare each dataset.

B. Ablation studies

Ablation studies are conducted in the base-to-novel setting. The average base accuracy, the average novel accuracy, and their harmonic mean are reported.

Insertion depth. The task-aware pre-context and learnable prompts are inserted into the encoders in a layer-wise manner or they are inserted into the input embeddings of all transformer blocks. We ablate the effect of the insertion depth by inserting the task-aware pre-context and learnable prompts into transformer blocks 1 to *insertion depth*. It is worth noting that (a) the depths for the textual and visual encoders are set to be equal by default, (b) the maximum depth is 12, and (c) the other ingredients are kept in this experiment. Table A1 shows the results. Our observations are as follows: (a) increasing the depth can improve the harmonic mean by a noticeable margin and (b) the biggest performance gains are achieved when the depth is 12. These observations demonstrate the effectiveness of inserting the task-aware pre-context and learnable prompts in a layer-wise manner.

Table A1. Ablation on the insertion depth. HM denotes the harmonic mean of the base and novel accuracies.

Insertion depth	Base	Novel	HM
1	81.69	72.95	77.07
3	83.17	74.60	78.65
6	84.19	76.68	80.26
9	84.93	76.26	80.36
12	**85.24**	**77.60**	**81.24**

Backbone structure. K-Means is conducted on the class embeddings that describe a downstream task and the clustering centers are treated as the backbone structure of the downstream task. We visualize class embeddings and corresponding clustering centers in Figure A1. It is worth noting that (a) the number of clustering centers is 5, (b) the dataset EuroSAT [4] only has 5 base or novel classes, and (c) PCA is used for dimensionality reduction. Our observations are as follows: (a) for the dataset EuroSAT, the class embeddings and clustering centers overlap due to their quantity being equal and (b) the backbone structure of class embeddings that describe a downstream task is captured. These observations demonstrate the effectiveness of capturing the backbone structure by task-aware clustering.

C. Discussion

This work explores the usage of clustering methods to preserve backbone structures of a downstream task, based on which the task-aware pre-context is generated and exploited to enhance the task-awareness of learnable prompts. Future research directions include developing better structure preservation techniques and enriching simple task descriptions, i.e., "dog" and "cat", with large language models to generate the comprehensive ones.

D. Source code

The source code is available at `https://github.com/FushengHao/TAC`. For more details, please see *README.md*.

(a) ImageNet base (b) ImageNet novel (c) Caltech101 base (d) Caltech101 novel

(e) OxfordPets base (f) OxfordPets novel (g) StanfordCars base (h) StanfordCars novel

(i) Flowers102 base (j) Flowers102 novel (k) Food101 base (l) Food101 novel

(m) FGVCAircraft base (n) FGVCAircraft novel (o) SUN397 base (p) SUN397 novel

(q) DTD base (r) DTD novel (s) EuroSAT base (t) EuroSAT novel

(u) UCF101 base (v) UCF101 novel

Figure A1. Visualization of class embeddings and corresponding clustering centers. The clustering centers capture the backbone structure of class embeddings that describe a downstream task. PCA is used for dimensionality reduction.

References

[1] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *ECCV*, pages 446–461, 2014. 1

[2] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *CVPR*, pages 3606–3613, 2014. 1

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 1

[4] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *J-STARS*, 12(7):2217–2226, 2019. 1

[5] 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*, pages 8340–8349, 2021. 1

[6] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *CVPR*, pages 15262–15271, 2021. 1

[7] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *CVPR*, pages 19113–19122, 2023. 1

[8] Muhammad Uzair Khattak, Syed Talal Wasim, Muzammal Naseer, Salman Khan, Ming-Hsuan Yang, and Fahad Shahbaz Khan. Self-regulating prompts: Foundational model adaptation without forgetting. In *ICCV*, pages 15190–15200, 2023. 1

[9] Jonathan Krause, Michael Stark, Jia Deng, and Fei-Fei Li. 3d object representations for fine-grained categorization. In *ICCV*, pages 554–561, 2013. 1

[10] Fei-Fei Li, Fergus Rob, and Perona Pietro. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *CVPR Workshop*, pages 178–178, 2004. 1

[11] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, pages 722–729. IEEE, 2008. 1

[12] Omkar M Parkhi, Andrea Vedaldi, Andrew Zisserman, and CV Jawahar. Cats and dogs. In *CVPR*, pages 3498–3505, 2012. 1

[13] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *ICML*, pages 5389–5400, 2019. 1

[14] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 1

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

[16] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *NeurIPS*, 2019. 1

[17] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, pages 3485–3492, 2010. 1

[18] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *CVPR*, pages 16816–16825, 2022. 1