Supplementary to Textual-based Class-aware Prompt tuning for Visual-Language Model

Hantao Yao[1],Rui Zhang[2], Changsheng Xu[1,3]

[1] State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation,Chinese Academy of Sciences, China

[2] State Key Lab of Processors, Institute of Computing Technology, Chinese Academy of Sciences, China;

[3] University of Chinese Academy of Sciences, China

hantao.yao@nlpr.ia.ac.cn; zhangrui@ict.ac.cn; csxu@nlpr.ia.ac.cn

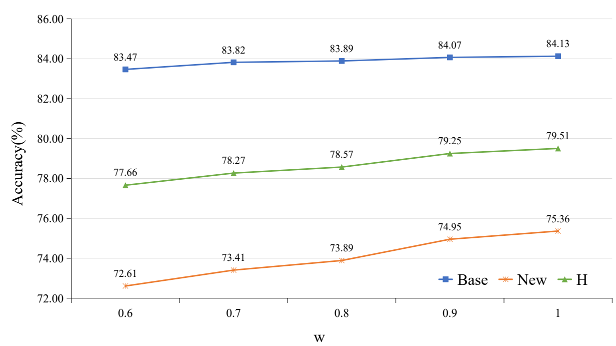

Figure 1. Effect of ω for prompt fusion.

1. Effect of Prompt Fusion

As shown in Eq (1)(Eq.(5) in the paper), we insert the obtained class-aware prompt into the mid-level textual tokens by replacing M-th textual tokens as the class-aware prompt,

$$\mathbf{F}'_l = [\mathbf{T}_1, \mathbf{T}_2, ..., \mathbf{T}_M, \mathbf{F}_{l,M+1}, \mathbf{F}_{l,M+2}, ..., \mathbf{F}_{l,N_t}]. \quad (1)$$

Note that Eq (1) discards the mid-level textual tokens $\hat{\mathbf{F}}_l = [\mathbf{F}_{l,1}, \mathbf{F}_{l,2}, ..., \mathbf{F}_{l,M}]$. We thus reformuate Eq (1) by fusing the class-aware prompt $\mathbf{T}$ and the discarded textual tokens $\hat{\mathbf{F}}_l$ with Eq. (2),

$$\mathbf{F}'_l = [\mathbf{T}'_1, \mathbf{T}'_2, ..., \mathbf{T}'_M, \mathbf{F}_{l,M+1}, \mathbf{F}_{l,M+2}, ..., \mathbf{F}_{l,N_t}], \quad (2)$$

where $\mathbf{T}'_m$ is the m-th fused textual tokens, which is the combination of the class-aware prompt $\mathbf{T}'_m$ and the mid-level textual tokens $\mathbf{F}_{l,m}$,

$$\mathbf{T}'_m = \omega\mathbf{T}'_m + (1-\omega)\mathbf{F}_{l,m}, \quad (3)$$

where ω is the weight.

We thus analyze the effect of ω for the TCP, and summarize the related results in Figure 1. As shown in Figure 1, a higher weight ω, a higher performance. Especially for the *New* performance on the unseen classes, an obvious performance improvement is obtained using higher ω. The reason is that a higher ω in Eq. (3) means that a fewer mid-level textual tokens biased to the training domain is considerred for the testing domain.

Table 1. Effect of class-aware prompt(CP)

	L_{kg}	CP	Base	New	H
CoOp			82.38	67.96	74.48
KgCoOp	√		80.73	73.6	77
TCP*		√	82.99	73.07	77.72
TCP	√	√	84.13	75.36	79.51

2. Class-aware Prompt *vs* regularize L_{kg}

The regularize term L_{kg} constrains the *output of TextEncoder*, while Class-aware Prompt explititly injects the class-related knowledge into the *middle layer of TextEncoder*. Moreover, class-aware prompt can explititly inject the class-level knowledge to increase the discrimiantive of textual-level classifier. As shown in Table 1, combining the Class-aware Promptand L_{kg} obtains a higher performance than merely using ones.

3. Comparison of training time and model complexity

As the additional learable parameters($\mathcal{P}$) in prompt tuning is far smaller than the fixed parameters($\mathcal{B}$), the inferrece time is major controlled by the backbone. Therefore, the methods with the same backbone(ViT-B/16) have the similar inference time(Tab. 2).

Table 2. Comparison of inference time and model complexity.

	KgCoOp	CoOp	PromptSRC	MaPLe	TCP
Total Parameters(M)	124.325	124.325	124.369	127.887	124.654
Fixed Parameters with ViT-B/16($\mathcal{B}$)(M)			124.323M		
Learnable Parameters ($\mathcal{P}$)(M)	0.002	0.002	0.046	3.564	0.331
GFlops	1547.42	1547.42	1547.82	1547.84	1547.59
Time(ms/batch)	124	124	124	124	124

Table 3. Comparison of domain generalization.

	ImageNet	ImageNet-V2	ImageNet-S	ImageNet-A	ImageNet-R	Avg.
CoCoOp	71.02	64.07	48.75	50.63	76.18	59.91
ProGrad	72.24	64.73	47.61	49.39	74.58	59.08
KgCoOp	71.2	64.1	48.97	50.69	76.7	60.12
MaPLe	70.72	64.07	49.15	50.9	76.98	60.27
DAPT	71.67	64.5	49.53	51.1	76.33	60.37
TCP	71.2	64.6	49.50	51.2	76.73	**60.51**

4. Domain Generalization

Domain Generalization aims to evaluate the generalization by evaluating the model on the target dataset having the same class but different data distribution from the source domain. Therefore, we conduct TCP on the few-shot ImageNets, and evaluate on the ImageNetV2, ImageNet-Sketch, ImageNet-A, and ImageNet-R. The related results are summarized in Table 3.

5. Datasets

Similar to existing CoOp-based methods, we conduct the evaluation on 11 datasets, *i.e.,* ImageNet [3], Caltech [4], OxfordPets [9], StanfordCars [6], Flowers [8], Food101 [1], FGVCAircraft [7], EuroSAT [5], UCF101 [10], DTD [2], and SUN397 [11]. As shown in Table 4, the type of datasets can be classified as: general object recognition, satellite image recognition, scene recognition, texture recognition, action recognition, and fine-grained object recognition.

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*, pages 446–461. Springer, 2014. 2, 3

[2] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 3606–3613. IEEE Computer Society, 2014. 2, 3

[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 Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society, 2009. 2, 3

[4] Li Fei-Fei, Robert Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Comput. Vis. Image Underst.*, 106(1):59–70, 2007. 2, 3

[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 J. Sel. Top. Appl. Earth Obs. Remote. Sens.*, 12(7):2217–2226, 2019. 2, 3

[6] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *2013 IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2013, Sydney, Australia, December 1-8, 2013*, pages 554–561. IEEE Computer Society, 2013. 2, 3

[7] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew B. Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *CoRR*, abs/1306.5151, 2013. 2, 3

[8] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Sixth Indian Conference on Computer Vision, Graphics & Image Processing, ICVGIP 2008, Bhubaneswar, India, 16-19 December 2008*, pages 722–729. IEEE Computer Society, 2008. 2, 3

[9] Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, pages 3498–3505. IEEE Computer Society, 2012. 2, 3

[10] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah.

Table 4. The detailed statistics of datasets used in our work.

Datasets	Classes	Training Size	Validation Size	Testing Size	Tasks
ImageNet [3]	1,000	1.28 M	N/A	50,000	General object recognition
Caltech [4]	100	4,128	1,649	2,465	General object recognition
EuroSAT [5]	10	13,500	5,400	8,100	Satellite image recognition
SUN397 [11]	397	15,880	3,970	19,850	Scene recognition
DTD [2]	47	2,820	1,128	1,692	Texture recognition
UCF101 [10]	101	7,639	1,808	3,783	Action recognition
FGVCAircraft [7]	100	3,334	3,333	3,333	Fine-grained aircraft recognition
OxfordPets [9]	37	2,944	736	3,669	Fine-grained pets recognition
StanfordCars [6]	196	6,509	1,635	8,041	Fine-grained car recognition
Flowers [8]	102	4,093	1,633	2,463	Fine-grained flowers recognition
Food101 [1]	101	50,500	20,200	30,300	Fine-grained food recognition

UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012. 2, 3

[11] Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*, pages 3485–3492. IEEE Computer Society, 2010. 2, 3