Supplementary Material

Retaining and Enhancing Pre-trained Knowledge in Vision-Language Models with Prompt Ensembling

Donggeun Kim[1,2*†] Yujin Jo[1*] Myungjoo Lee[1*] Taesup Kim[1‡]
[1]Graduate School of Data Science, Seoul National University [2]Nota Inc.

Table 1. Parameter comparison with other models

Method	Params	Params % CLIP	H
CoOp	2048	0.002	71.66
CoCoOp	35360	0.03	75.83
RPO	30720	0.02	77.78
MaPLe	3.55M	2.85	78.55
PromptSRC	46K	0.04	79.97
GPE	30720	0.02	79.24

The section below includes additional information, a comparison of parameter efficiency, and additional ablation studies of GPE.

1. Parameter Efficiency Comparison with Different Prompting Methods

Table 1 represents the number of trainable parameters and the harmonic mean on base-to-novel generalization in comparison with CoOp [6], CoCoOp [5], RPO [3], MaPLe [1], PromptSRC [2], and GPE. As shown, GPE outperforms other methods updating a similar number of parameters with a remarkable performance difference. Even when compared to MaPLe, which has a significantly larger number of learnable prompts, GPE achieves superior performance with considerably fewer parameters.

2. Additional Ablation Studies

Pre-softmax vs. Post-softmax for Inference Our investigation reveals that making predictions by the traditional ensemble method, which averages softmax-transformed logits, yields better results. In comparison with the pre-softmax approach outlined in Table 2, where logits are averaged before softmax for inference, our default method exhibits enhanced performance. This emphasizes the importance of considering the order of operations in ensemble methods, particularly when optimizing model performance.

Covariance Regularization Effect When applying a loss function that drives each prompt away from the centroid of all prompts, as suggested in C-TPT [4], instead of covariance loss, the performance dropped to 78.04%, as shown in Table 2.

[*]Equal contribution.
[†]Work done at Seoul National University.
[‡]Corresponding author.

Table 2. Ablation study on GPE methods

Method	Base	Novel	H
GPE	**83.26**	**75.92**	**79.24**
GPE w/ pre-softmax inference	83.43	75.68	79.17
GPE w/ centroid loss	81.62	74.76	78.04

References

[1] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19113–19122, 2023. 1

[2] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15190–15200, 2023. 1

[3] Dongjun Lee, Seokwon Song, Jihee Suh, Joonmyeong Choi, Sanghyeok Lee, and Hyunwoo J Kim. Read-only prompt optimization for vision-language few-shot learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1401–1411, 2023. 1

[4] Hee Suk Yoon, Eunseop Yoon, Joshua Tian Jin Tee, Mark Hasegawa-Johnson, Yingzhen Li, and Chang D Yoo. C-tpt: Calibrated test-time prompt tuning for vision-language models via text feature dispersion. *arXiv preprint arXiv:2403.14119*, 2024. 1

[5] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 1

[6] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 1