Prompt Learning via Meta-Regularization (Supplement)

Jinyoung Park, Juyeon Ko, Hyunwoo J. Kim*
Department of Computer Science and Engineering, Korea University
{lpmn678, juyon98, hyunwoojkim}@korea.ac.kr

In this supplement, we provide the implementation details (Section A) and additional experimental results (Section B).

A. Implementation details

In this section, we provide the implementation details of our work. We implement our ProMetaR (**Pro**mpt learning via **Meta R**egularization) using Pytorch [6] and Dassl [11, 12], which is a library designed for domain adaptation and generalization. Following previous prompt learning methods [4, 5, 13], we use CLIP-ViT-B/16 as the pretrained backbone model [7] and four soft prompting tokens for each modality. Following other works [1, 5, 7], we utilize an ensemble of text prompts for the textual regularizer. For the base prompt learning method, we use Independent Vision-Language Prompting as a base prompt learning method that optimizes hierarchical prompts on both image and text modalities [4]. The learning rate is set to 0.0025, and the prompts are optimized with SGD optimizer for all experiments. For the base-to-new generalization settings, we train the model for 15 epochs. For domain generalization and cross-dataset transfer settings, we train the models for 6 epochs. In all experiments, we evaluate the performance of the methods in three independent runs (seed 1, 2, and 3) and report average performance following previous prompt learning approaches [4, 5, 13].

Evaluation metrics. In all experiments, we report top-1 accuracy for each dataset. In base-to-novel generalization, the top-1 accuracy is measured on base classes and new classes, respectively. We calculate the harmonic mean (H) between the base and new class accuracy to show the generalization trade-off [9]. In domain generalization, and cross-dataset evaluation settings, we measure top-1 accuracy on the test set of each dataset with the split provided by CoOp [14] following other prompt optimization works.

*is the corresponding author.

B. Additional experiments

In this section, we provide the results of the additional experiments including cross-dataset settings and more analysis.

B.1. Cross-dataset

We also measure the performance of the proposed method in the cross-dataset transfer setting to explore the *task* generalization ability of ProMetaR in Table 1. In cross-dataset transfer setting, we train our ProMetaR on ImageNet [2] as a source dataset and evaluate it on other 11 *unseen* datasets such as Caltech101, OxfordPets, StanfordCars, Flowers102, Food101, FGVC Aircraft, Sun397, DTD, EuroSAT, and UCF101 following other works. Please note that the model cannot access the *unseen* datasets during the training phase.

For a fair comparison, we exclude UNIGRAM since it employs a large scale of extra datasets to pre-train the learnable prompts. From the table, ProMetaR successfully generalizes on out-of-domain datasets, achieving the best performance on 7 out of 10 datasets compared to other baselines. This result indicates that our ProMetaR improves the *task* generalization ability of the existing prompting methods and robustness against domain shifts.

B.2. More analysis

Comparison of ProMetaR with the generalization methods. We examine the efficacy of ProMetaR by comparing ours with data augmentation methods: Mixup [10] and Manifold Mixup [8] and common generalization methods based on the weight averaging: exponential moving average (EMA) and stochastic weight averaging (SWA) [3] by applying them to the base prompt learning method, IVLP. The results are reported in Table 2. Mixup slightly improves the performance on base classes with an accuracy gain of 0.39%, but it shows the performance degradation on the new classes. Similarly, Manifold Mixup decreases the performance on new classes with the performance gain on base classes. These results indicate that conventional data augmentation helps improve the performance on base classes (*traditional* generalization), but it still suffers from

	Source	Target									
	ImageNet	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	Aircraft	SUN397	DTD	EuroSAT	UCF101
CoOp	71.51	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	46.39	66.55
CoCoOp	71.02	**94.43**	90.14	65.32	71.88	86.06	22.94	67.36	45.73	45.37	68.21
MaPLe	70.72	93.53	90.49	65.57	**72.23**	86.20	24.74	67.01	46.49	**48.06**	68.69
PromptSRC	71.27	93.60	90.25	65.70	70.25	86.15	23.90	67.10	46.87	45.50	68.75
UNIGRAM	71.65	94.67	90.83	66.78	73.12	86.69	25.27	67.97	48.06	52.63	71.03
ProMetaR	**71.29**	93.74	**90.59**	**65.83**	71.13	**86.39**	**24.78**	**67.41**	**47.08**	45.02	**69.50**

Table 1. Performance comparison on the cross-dataset transfer setting.

Methods	**Base**	**New**	**H**
IVLP (Base)	82.51	73.36	77.66
Mixup	82.90 (+0.39)	71.45 (-1.91)	76.75 (-0.91)
Manifold Mixup	83.57 (+1.06)	73.19 (-0.17)	78.04 (+0.38)
EMA	82.30 (-0.21)	74.15 (+0.79)	78.01 (+0.35)
SWA	83.65 (+1.14)	73.14 (-0.22)	78.04 (+0.38)
ProMetaR **(Ours)**	**84.39 (+1.88)**	**76.93 (+3.57)**	**80.49 (+2.83)**

Table 2. Performance comparison of ProMetaR with the domain generalization methods on the base-to-new generalization setting. Results are averaged over 11 datasets. H refers to harmonic mean.

Methods	**Base**	**New**	**H**
No TaskAug	84.27	75.06	79.40
TaskAug: Input Mixup	84.26	76.10	79.97
TaskAug: Manifold Mixup **(Ours)**	**84.39**	**76.93**	**80.49**

Table 3. Effect of our proposed meta-regularization. Results are averaged over 11 datasets. H refers to harmonic mean.

the task overfitting problem in existing prompt learning methods to generalize on the new classes (*task* generalization). EMA enhances new class accuracy by +0.79%, at little expense of base class accuracy. Meanwhile, SWA improves performance on base classes with an improvement of +1.14%, but the average accuracy on new classes slightly decreases. We observe that our ProMetaR significantly outperforms both domain augmentation and generalization methods by a large margin.

Task augmentation. In Table 3, we measure the performance of the model without using task augmentation (No TaskAug), with the input Mixup [10] for a task augmentation and our ProMetar that uses Manifold Mixup for the task augmentation. Compared to No TaskAug, task augmentation improves the performance on new classes without the loss of the performance on the base classes. This demonstrates that using task augmentation alleviates the meta-overfitting issue by generating various virtual augmented tasks. In addition, the task augmentation with the manifold mixup shows better performance than the input mixup with a performance gain of 0.83% on new classes.

References

[1] James Urquhart Allingham, Jie Ren, Michael W Dusenberry, Xiuye Gu, Yin Cui, Dustin Tran, Jeremiah Zhe Liu, and Balaji Lakshminarayanan. A simple zero-shot prompt weighting technique to improve prompt ensembling in text-image models. In *ICML*, 2023. 1

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

[3] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. In *UAI*, 2018. 1

[4] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad

Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. In *CVPR*, 2023. 1

[5] 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*, 2023. 1

[6] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *ICLRW*, 2017. 1

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1

[8] Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Aaron Courville, Ioannis Mitliagkas, and Yoshua Bengio. Manifold mixup: learning better representations by interpolating hidden states. In *ICML*, 2019. 1

[9] Yongqin Xian, Bernt Schiele, and Zeynep Akata. Zero-shot learning-the good, the bad and the ugly. In *CVPR*, 2017. 1

[10] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 1, 2

[11] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain adaptive ensemble learning. *TIP*, 30:8008–8018, 2021. 1

[12] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *TPAMI*, 2022. 1

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

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