Supplemental Material for Texts as Images in Prompt Tuning for Multi-Label Image Recognition

Zixian Guo[1,2*] Bowen Dong[1] Zhilong Ji[2] Jinfeng Bai[2] Yiwen Guo[4] Wangmeng Zuo[1,2(✉)]
[1]Harbin Institute of Technology [2]Tomorrow Advancing Life [3]Pazhou Lab, Guangzhou [4]Independent Researcher
zixian_guo@foxmail.com cndongsky@gmail.com zhilongji@hotmail.com
jfbai.bit@gmail.com guoyiwen89@gmail.com wmzuo@hit.edu.cn

A. Overview

This material provides more information on our TaI-DPT and experimental results. The supplementary material is organized as follows. In Sec. B, we present more details about our prepared text data used for training. In Sec. C, we display more ablation studies on the training loss, different VL pre-trained models, texts v.s. images for prompting and the coefficients used in the prompt ensemble. In Sec. D, we discuss the connection and distinction between our method and the dual-prompt design proposed in [10].

B. More Details about Text Descriptions

B.1. Noun Filtration

To extract the category labels from texts exhaustively, we construct synonym dictionaries for classes involved in VOC2007 [4], MS-COCO [8], and NUS-WIDE [3] by gathering the expressions of the classes from different sources. We use the WordNet [5] interface provided by [2] to get a relatively comprehensive list of synonyms and then manually select words with specific meanings for inclusion in the synonym dictionary. In addition, we also collect expressions for categories from standard online dictionaries. Besides, some words exist in the corpus in simple and compound forms, like "cellphone" and "cell phone", and we prioritize compound word matches. Since the 80 categories of MS-COCO [8] cover the categories of VOC2012 [4], for these two datasets, we filtered the captions from MS-COCO using the same synonym dictionary (shown in "synonyms_COCO.txt") to obtain the texts and labels as the training data. For NUS-WIDE [3], we introduce localized narratives from OpenImages [7], which have a broader range of content, to cover all the concepts in NUS-WIDE. The synonym dictionary for NUS-WIDE is shown in "synonyms_NUSWIDE.txt".

*This work was done when Zixian Guo was a research intern at TAL.

Table A. Comparison of the results when training TaI-DPT with different learning objectives. Ranking loss (RL) [6] is a properer and more flexible way to guide the learning of prompts.

Loss	VOC2007	MS-COCO	NUSWIDE
BCE	84.9	59.0	40.5
ASL [1]	84.6	56.9	36.0
RL [6]	**88.3**	**65.1**	**46.5**

B.2. Hand-craft Prompt Templates

Using the noun filtration strategy above, we end up with 66,087, 100,543, and 456,759 pieces of text for VOC2007, MS-COCO, and NUS-WIDE, respectively. Even for some common categories, the amount of texts is relatively sufficient, but we still find that there are few occurrences of certain categories in the texts. Especially for objects that are not prominent on which the text descriptions tended not to focus. So to process these categories better, we also added the hand-crafted prompt templates for each class as training data. The used templates are listed in "prompt_templates.txt".

C. More Ablation Studies

C.1. Loss Function

As explained in Sec. 3.4 of our main paper, we discussed the loss function used to train our TaI-DPT. Here, we provide the results on the three datasets when training with common binary cross-entropy loss (BCE), asymmetric loss (ASL) [1], and ranking loss (RL) [6]. Formally, the binary cross-entropy loss is defined as:

$$\mathcal{L} = \mathrm{BCE}(\boldsymbol{p}, \boldsymbol{y}) + \mathrm{BCE}(\boldsymbol{p}', \boldsymbol{y}),$$
$$\mathrm{BCE}(\boldsymbol{q}, \boldsymbol{y}) = -\frac{1}{C}\sum_{i=1}^{C}[\boldsymbol{y}_i \cdot \log \boldsymbol{q}_i + (1-\boldsymbol{y}_i)\cdot \log(1-\boldsymbol{q}_i)] \quad (1)$$

Table B. Results of other pre-trained VL models.

VL model	ZeroShot		TaI-DPT	
	VOC2007	MS-COCO	VOC2007	MS-COCO
OpenCLIP (ViT-B/32, LAION-2B)	80.5	52.5	87.7	64.2
DeCLIP (ResNet50, 88M)	77.6	43.8	80.3	46.2

Table C. The results of training the double-grained prompt with text data and labeled images on VOC2007. Our TaI-DPT can learn effective prompts in the zero-shot setting.

Method	DPT	ZSCLIP	TaI	Image
VOC2007	✗	76.2	86.0	90.0
	✓	77.3	88.3	93.9

where $\boldsymbol{p}$ and $\boldsymbol{p}'$ are global and local classification score. And the asymmetric loss is defined as:

$$\begin{aligned} \mathcal{L} &= \mathrm{ASL}(\boldsymbol{p}, \boldsymbol{y}) + \mathrm{ASL}(\boldsymbol{p}', \boldsymbol{y}), \\ \mathrm{ASL}(\boldsymbol{q}, \boldsymbol{y}) &= -\frac{1}{C}\sum_{i=1}^{C}[\boldsymbol{y}_i \cdot k_+ + (1-\boldsymbol{y}_i)\cdot k_-], \\ k_+ &= (1-\boldsymbol{q}_i)^{\gamma_+}\log \boldsymbol{q}_i, \\ k_- &= (\boldsymbol{q}_i^m)^{\gamma_-}\log\,(1-\boldsymbol{q}_i^m) \end{aligned} \tag{2}$$

where $\boldsymbol{q}^m = \max(\boldsymbol{q} - m, 0)$ and hyperparameters γ_+, γ_- and m are set as 1, 2 and 0.05, respectively, according to [10]. The training results with different losses are shown in Table A.

C.2. Effects of various VL models

From Table B, our approach consistently improves on different VL models trained with different data sources. Pre-trained with much fewer data, DeCLIP performs weaker in zero-shot recognition and adaptation.

C.3. Texts v.s. Images for Prompting

To directly compare the difference between prompting with texts and prompting with images, we train our double-grained prompt with images (I-DPT) from trainval set and compare it with TaI-DPT on the test set of VOC2007 [4]. The results are shown in Table C. It's obvious that we can learn the prompts well with sufficient labeled images, improving the mAP of zero-shot CLIP from 77.3 to 93.9. However, when no image data is available, our TaI-DPT can reach 88.3 mAP, demonstrating the effectiveness of our zero-shot prompt tuning scheme.

C.4. Summation Coefficient in Prompt Ensemble

As illustrated in Sec. 3.5 of our main paper, our TaI-DPT can easily combine with existing prompting methods learned with images and yield complementary improvements. Here, we explore the coefficient used to fuse the classification score produced by different models. For example, let $\boldsymbol{p}_1$ denote the score provided by CoOp [11] and $\boldsymbol{p}_2$ denotes the score yielded by our TaI-DPT. The merged score is obtained by weighted summation $\boldsymbol{p} = \lambda \cdot \boldsymbol{p}_1 + (1-\lambda)\cdot \boldsymbol{p}_2$.

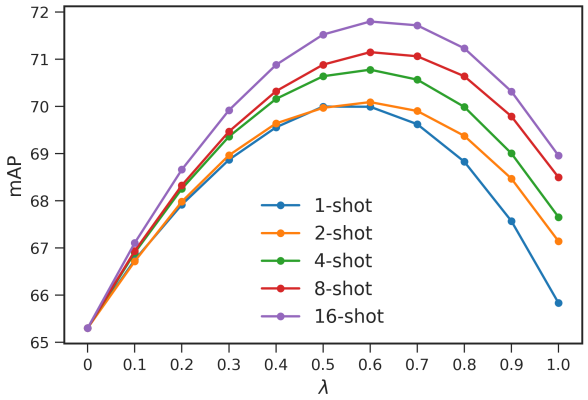

Figure A. Relation between ensemble performance on MS-COCO and summation coefficient.

Table D. The results of using **hand-crafted** positive and negative templates during the zero-shot inference of CLIP [9]. Despite containing a completely opposite meaning, the negative linguistic inputs still achieve considerable accuracy.

Template	VOC2007	MS-COCO	NUSWIDE
Pos.	76.2	47.3	36.4
Neg.	66.2	41.8	24.3

From Fig. A, we can see the change of mAP of $\boldsymbol{p}$ relative to coefficient λ. So we set $\lambda = 0.6$ for the ensemble of TaI-DPT and CoOp-DPT learned from few-shot samples, which gives better results in various few-shot settings. Similarly, we set $\lambda = 0.9$ when combining our TaI-DPT with DualCoOp [10] when partially annotated images are available.

D. Comparison with DualCoOp

As the first approach to adapt pre-trained CLIP [9] to multi-label recognition tasks, DualCoOp [10] proposes to use a pair of contrastive positive and negative prompts to generate binary classification probability for each class. However, the negative prompt may not be a property way to adapt CLIP. In Table D, we show zero-shot recognition results of CLIP [9] with hand-crafted positive and negative templates. We use a positive template, "a photo of a [CLASS]" and a negative template, "a photo without [CLASS]". It seems that the negative prompt is dominated by the [CLASS] token and still gives rise to considerable recognition accuracy as the positive prompt does, which can make it reluctant to analyze the effect of a negative prompt.

But for our proposed double-grained prompt tuning (DPT), the two prompts are all positive and focus on global and local features separately. Intuitively, the global prompt

can be seen as a hand-crafted prompt like "a photo of a [CLASS]", and the local prompt can be seen as "a cropped photo of a [CLASS]". The two positive prompts can be learned flexibly with ranking loss [6], without relying on each other to produce a classification score for each class.

Besides, DualCoOp [10] uses all images from the training set with partial labels to learn the prompts. Our TaI-DPT advocates using descriptive texts as an alternative when there is no image data, and the pseudo-label for each text derived with noun filtration can be regarded as incomplete categorical labels. As such, our prepared text data is somewhat homogeneous with the partial-labeled image data, which leads to gentle improvements when combining our method with DualCoOp. However, in the case of few-shot image samples available, our TaI-DPT brings considerable enhancements by ensemble with the few-shot approach like CoOp [11] as shown in Fig. A.

References

[1] Emanuel Ben-Baruch, Tal Ridnik, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification, 2021. 1

[2] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009. 1

[3] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. Nus-wide: a real-world web image database from national university of singapore. In *Proceedings of the ACM international conference on image and video retrieval*, pages 1–9, 2009. 1

[4] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 1, 2

[5] Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998. 1

[6] Yunchao Gong, Yangqing Jia, Thomas Leung, Alexander Toshev, and Sergey Ioffe. Deep convolutional ranking for multilabel image annotation. *arXiv preprint arXiv:1312.4894*, 2013. 1, 3

[7] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Shahab Kamali, Matteo Malloci, Jordi Pont-Tuset, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://storage.googleapis.com/openimages/web/index.html*, 2017. 1

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 1

[9] 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 *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2

[10] Ximeng Sun, Ping Hu, and Kate Saenko. Dualcoop: Fast adaptation to multi-label recognition with limited annotations. *arXiv preprint arXiv:2206.09541*, 2022. 1, 2, 3

[11] 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. 2, 3