Learning Through Consistency for Prompt Tuning

Shuvendu Roy, Ali Etemad
Dept. ECE and Ingenuity Labs Research Institute
Queen's University, Kingston, Canada
{shuvendu.roy, ali.etemad}@queensu.ca

Abstract

We propose Consistency-guided Prompt learning (CoPrompt), a new fine-tuning method for vision-language models that addresses the challenge of improving the generalization capability of large foundation models while fine-tuning them on downstream tasks in a few-shot setting. The basic idea of CoPrompt is to enforce a consistency constraint in the prediction of the trainable and pre-trained models to prevent overfitting on the downstream task. Additionally, we introduce the following two components into our consistency constraint to further boost the performance: enforcing consistency on two perturbed inputs and combining two dominant paradigms of tuning, prompting and adapter. Enforcing consistency on perturbed input further regularizes the consistency constraint, effectively improving generalization, while tuning additional parameters with prompting and adapters improves the performance on downstream tasks. Experiments show that CoPrompt outperforms existing methods on various evaluation suites, including base-to-novel generalization, domain generalization, and cross-dataset evaluation tasks. On the generalization task, CoPrompt improves the state-of-the-art by 2.09% on the zero-shot task and 1.93% on the harmonic mean over 11 recognition datasets. Detailed ablation studies show the effectiveness of each of the components in CoPrompt.

1 Introduction

Vision-language foundation models (e.g., CLIP [1]) that are trained on large-scale datasets of image-text pairs have demonstrated excellent generalization capabilities. However, the sheer size of these models can make it challenging to fine-tune them for downstream tasks, especially for small downstream tasks (e.g., few-shot learning), while preserving their ability to generalize [2, 3]. Despite recent advancements in few-shot fine-tuning, it is still a challenge to maintain the generalization capability of the original models such as CLIP, let alone improve them. In fact, it has been shown that improvements in few-shot performance often result in a drop in zero-shot capabilities (e.g., CoOp [2]). This is caused by severe overfitting on newly introduced parameters during few-shot fine-tuning, resulting in a significant deviation from the foundation model's original behaviour.

In this work, we propose Consistency-guided Prompt learning (CoPrompt), a new fine-tuning method for vision-language models that reduces the overfitting problem by preventing the trainable model's embeddings from deviating too far from the pre-trained model's embedding. More specifically, we enforce a consistency constraint on both the language and image branches between the trainable and pre-trained models. Furthermore, we introduce two additional components to improve the proposed consistency constraint. First, we enforce consistency on two perturbed inputs instead of the same input to further regularize the consistency constraint, effectively improving generalization. On the 'text' branch, we use a pre-trained large language model (LLM), GPT [4], to generate a more detailed and descriptive sentence from an input prompt text of a generic format (a photo of a 'class'). We then enforce consistency between the learnable and pre-trained text encoder on the representations of these two sentences. On the image branch, we apply augmentations on an input image to generate two perturbed images. Second, we add more trainable parameters to increase the model's learning

R0-FoMo: Workshop on Robustness of Few-shot and Zero-shot Learning in Foundation Models at NeurIPS 2023.

capacity on the new task. To this end, we combine the two dominant paradigms of tuning, prompting [5], and adapters [6]. By combining prompt and adapter, we can tune more parameters to improve the performance on the new task while the consistency constraint maintains or possibly improves the model's zero-shot capability.

Extensive experiments on three common evaluation settings demonstrate the strong performances of CoPrompt. In the base-to-novel generalization task, CoPrompt outperforms existing methods on 11 benchmark datasets, achieving a 2.09% improvement on novel classes and a 1.93% improvement on the harmonic mean over MaPLe, the previous SOTA. Importantly, our improvements do not come at the expense of the performance of the base class, which also achieves a 1.72% improvement over MaPLe. Additionally, CoPrompt achieves considerable improvements over existing methods on cross-dataset evaluation and domain generalization. An extensive ablation study confirms the importance of each component of the proposed method. In summary, this paper makes the following contributions: (**1**) We propose a consistency-enforced fine-tuning method for large foundation models that enables learning a new task from a few samples without losing its zero-shot generalizability. (**2**) Our method incorporates the knowledge of a pre-trained LLM with consistency constraints on the text branch and data augmentations on the image branch to improve the generalization further. (**3**) Our method combines the two strong paradigms of tuning foundation models, prompting and adapter, into a single framework to improve performance on new tasks. (**4**) We set a new state-of-the-art for a range of evaluation suites, including base-to-novel generalization, cross-dataset recognition, and domain generalization.

2 Related Work

Recent developments in vision-language models, such as CLIP [1], ALIGN [7], LiT [8], FILIP [9], and Florence [10], have exhibited impressive performance in various vision tasks, including few-shot and zero-shot learning. However, the enormous size of these models makes it challenging to fine-tune them without losing their generalization. The two commonly used approaches for using a pre-trained model for a downstream task are (a) full fine-tuning and (b) linear probing. However, neither of these methods performs well for foundation models. Full fine-tuning results in a loss of generalization, while linear probing often leads to poor performance on downstream tasks [5]. Consequently, many recent studies have focused on adapting large foundation models on downstream tasks without changing the pre-trained weights [2]. Existing works in this direction can be categorized into two main groups: Prompting [3, 5] and Adapter [6].

Prompts are typically instructions in the form of text that guide the downstream task. They can either be manually crafted for a specific task or learned automatically. The latter method is called prompt tuning, which was initially introduced in [11, 12, 13]. In this context, CoOp [2] introduced a set of continuous vectors into the language branch's input, which is optimized with the final loss. However, this approach demonstrated poor performance on unseen classes, indicating poor generalization on zero-shot tasks. CoCoOp [3] improved CoOp's zero-shot performance by explicitly conditioning on the image inputs. ProGrad [14] only updated the prompts where the gradients aligned with the original prompt's defined general knowledge. Prompting techniques have also been utilized for dense prediction tasks [15]. While the earlier works on prompting added prompts only to the text input, some recent works have also explored prompting on the image inputs [16]. Later, MaPLe took a multi-modal approach that used prompting on both image and text inputs. This method explicitly ensured mutual synergy between the language and vision prompts to discourage learning from unimodal data. Another method for tuning foundation models is Adapters. This approach introduces learnable parameters to one or multiple layers of the pre-trained model to transform features [6]. Adapters are typically added to the upper layers of the network, which can be seen as a learnable transformation module for the pre-trained model. Adapters have also been studied in vision-only models, including dense prediction tasks [17].

3 Method

CoPrompt tackles the issue of reduced generalization due to overfitting on the downstream task, by implementing a consistency constraint that ensures the text and image embeddings produced by the trainable model (tunable prompt parameters in both the image and text branches) are not significantly different from those generated by the pre-trained CLIP.To further impose regularization in the consistency constraint, we utilize perturbations to the input for the trainable and pre-trained models. On the language branch, a pre-trained LLM is used to generate more descriptive sentences from

the template text input, while on the image branch, we use augmentations. In addition, CoPrompt includes additional trainable parameters by adding adapters on the image and text branches to enhance performance on new downstream tasks. In this work, we build upon the prompting concept of MaPLe [5], which utilizes a coupling function F, to condition the image prompt on the text prompt. An overview of CoPrompt is illustrated in Figure 1.

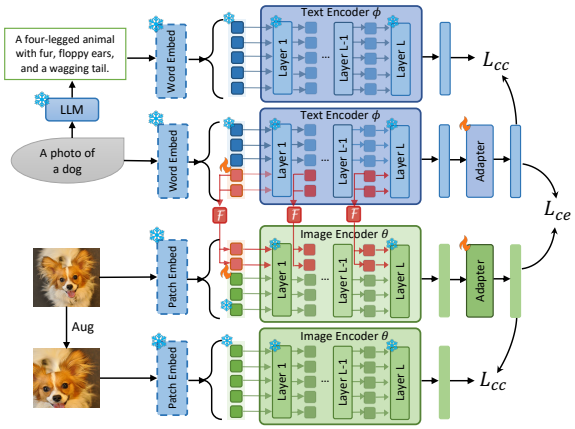

Figure 1: Overview of the proposed CoPrompt.

Consistency constraint. We use cosine distance as the consistency constraint between the embeddings of the pre-trained encoder and the learnable encoder. However, other similar criteria, like Euclidean distance, can also be used as a constraint. We empirically observe that cosine distance as the consistency constraint yields the best performance. This constraint is applied on image and text branches. We can denote the consistency constraint as:

$$\mathcal{L}_{cc} = 2 - \frac{w_y \cdot \phi(t_y)}{||w_y||\ ||\phi(t_y)||} - \frac{z \cdot \theta(i)}{||z||\ ||\theta(i)||}. \tag{1}$$

Input perturbation. Given the template text 'a photo of a [category]', we use a pre-trained LLM, GPT(ϕ_{GPT}), to generate a more descriptive sentence as $s_k = \phi_{GPT}$('a photo of a [category]$_k$'). For this, we follow the training setup of KgCoOP [18]. But unlike KgCoOp, we generate a single sentence on the fly rather than generating a pre-defined number of sentences and averaging their embedding. On the image branch, we use an augmentation module δ to generate perturbed image $x' = \delta(x)$. Now we enforce the consistency between the embedding of the perturbed input to the pre-trained model and the learnable model as:

$$\mathcal{L}_{cc} = 2 - \frac{\phi(s_y) \cdot \phi(t_y)}{||\phi(s_y)||\ ||\phi(t_y)||} - \frac{\theta(x') \cdot \theta(i)}{||\theta(x')||\ ||\theta(i)||}. \tag{2}$$

Adapters. We incorporate more trainable parameters for better adaptation to the new task. We do so by adding adapters [6] to both the vision and language branches. Adapters are trainable parameters that are added on top of the encoder to transform the embedding vector. Following [6], we define our adapter as two linear layers with non-linearity in between. But unlike [6], we do not restrict the adapter only to the text branch; rather use it on both. Let ϕ^a be the text adapter that takes a text embedding w_k as input and transforms it as $\phi^a(w_k)$. Similarly, θ^a is the image adapter. The proposed consistency constraint loss can be represented as:

$$\mathcal{L}_{cc} = 2 - \frac{\phi(s_y) \cdot \phi^a(\phi(t_y))}{||\phi(s_y)||\ ||\phi^a(\phi(t_y))||} - \frac{\theta(x') \cdot \theta^a(\theta(i))}{||\theta(x')||\ ||\theta^a(\theta(i))||}. \tag{3}$$

Final loss. The proposed consistency constraint loss is combined with a supervised loss to form the final loss. We represent the supervised loss as: $\mathcal{L}_{ce} = -log\frac{\exp(sim(z,w_y)/\tau)}{\sum_{k=1}^{C}\exp(sim(z,w_k)/\tau)}$. Adding both losses with a balancing factor λ, we get the final loss function of CoPrompt is: $\mathcal{L} = \mathcal{L}_{ce} + \lambda\mathcal{L}_{cc}$.

4 Experiments

Base to Novel Generalization. We present the results of our proposed method on the base-to-novel generalization task and compare with prior works in Table 2. We have highlighted the best results in bold and marked the improvement over the previous state-of-the-art (MaPLe) in blue. As we see from the average over all datasets (Table 1), CoPrompt outperforms all existing methods for both novel and base categories by a large margin. Our method demonstrates strong zero-shot generalization, with an improvement of 2.09% on novel categories, increasing the SOTA from 75.14% to 77.23%. Apart from MaPLe, no existing method outperformed the pre-trained CLIP (which was not fine-tuned), indicating the difficulty

Table 1: Performance on base-to-novel generalization

	Base	Novel	HM
CLIP	69.34	74.22	71.70
CoOp	82.69	63.22	71.66
Co-CoOp	80.47	71.69	75.83
ProGrad	82.48	70.75	76.16
KgCoOp	80.73	73.60	77.00
MaPLe	82.28	75.14	78.55
CoPrompt	**84.00**	**77.23**	**80.48**
	+1.72	+2.09	+1.93

of maintaining zero-shot performance while learning a new task in a few-shot setting. Along with a large improvement in zero-shot performance, CoPormpt also improves the few-shot performance by 1.72% on base categories. This confirms that improvement in zero-shot performance does not come at the cost of few-shot performance or vice versa. In fact, CoPrompt is the first method to beat all existing methods in both base and novel categories since the introduction of prompt tuning of VLMs in CoOp. On a harmonic mean, CoPrompt provides a 1.93% improvement over existing methods.

Cross-dataset Evaluation. In Table 2, we present the results for cross-dataset evaluation. Here, the model is fine-tuned on a source dataset (ImageNet) and evaluated on target datasets in a zero-shot manner. As we see, CoPrompt shows improvements on 9 out of 10 datasets. Overall, CoPromt provides an average accuracy of 67.0%, which is 0.70% higher than the previous state-of-the-art, MaPLe. We obtain the highest improvement of 3.84% over MaPLe on the EuroSAT dataset.

Table 2: Performance of CoPrompt on cross-dataset evaluation.

	Caltech	Pets	Cars	Flowers	Food	Aircraft	SUN397	DTD	EuroSAT	UCF	*Ave.*
CoOp	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	46.39	66.55	63.88
Co-CoOp	94.43	90.14	65.32	71.88	86.06	22.94	67.36	45.73	45.37	68.21	65.74
MaPLe	93.53	90.49	65.57	72.23	86.20	**24.74**	67.01	46.49	48.06	68.69	66.30
CoPrompt	**94.50**	**90.73**	**65.67**	**72.30**	**86.43**	24.00	**67.57**	**47.07**	**51.90**	**69.73**	**67.00**

Domain Generalization. We present the results for domain generalization in Table 3. Here, the original ImageNet dataset is used as the source dataset to fine-tune the model. The model is then tested on four other variants of ImageNet that come from different distributions. CoPrompt shows strong domain generalization by outperforming MaPLe on all datasets except for ImageNet-A. It achieves a new state-of-the-art average accuracy of 60.43% on this task.

Table 3: Performance on domain generalization.

	Source	**Target**				
	ImNet	ImNetV2	ImNetS	ImNetA	ImNetR	Ave.
CLIP	66.73	60.83	46.15	47.77	73.96	57.17
UPT	**72.63**	64.35	48.66	50.66	76.24	59.98
CoOp	71.51	64.20	47.99	49.71	75.21	59.28
Co-CoOp	71.02	64.07	48.75	50.63	76.18	59.90
ProGrad	72.24	64.73	47.61	49.39	74.58	59.07
KgCoOp	71.20	64.10	48.97	50.69	76.70	60.11
MaPLe	70.72	64.07	49.15	**50.90**	76.98	60.26
CoPrompt	70.80	**64.25**	**49.43**	50.50	**77.51**	**60.42**

Ablation Study. In this section, we present an ablation study by removing different components of the proposed method to understand the importance of each of them. We show the results of these experiments in Table 4. For reference, in the first row of the table, we present the final performance of CoPrompt, which has a harmonic mean of 80.48%. In the first ablation experiment, we eliminate the adapters from CoPrompt, leading to an accuracy of 80.02% (a performance drop of 0.46%). This highlights the importance of adapters in CoPrompt. Next, we remove the input perturbations, effectively enforcing consistency between the trainable and pre-trained encoder for the same image and text input. This results in an accuracy of 79.56%, which is a 0.92% drop in performance, suggesting the high importance of input perturbations in CoPrompt. Then, we remove both the input perturbations and the adapters, which results in an average accuracy of 79.50%. This shows that utilizing the consistency constraint alone provides a 0.95% improvement over removing all three components (as shown in the last row of the table). In the last study, we remove the consistency constraint along with the input perturbations, effectively training the adapters and prompts without enforcing consistency. Surprisingly, this results in an accuracy of 78.45%, even lower than when all three components are removed. Additional experiments can be found in the full version of our paper at arxiv.org/abs/2306.01195.

Table 4: Ablation Study

Cons.	In. Pert.	Adp.	Accuracy
✓	✓	✓	80.48
✓	✓	✗	80.02
✓	✗	✓	79.56
✓	✗	✗	79.50
✗	✗	✓	78.45
✗	✗	✗	78.55

5 Conclusion

We present a novel tuning method for large vision-language foundation models that enhances their performance in downstream tasks and also improve zero-shot generalization. CoPrompt is a carefully designed method with three important components that reduce the overfitting problem during fine-tuning. Through extensive evaluations across three different tasks, CoPrompt demonstrated its effectiveness in few-shot learning, zero-shot learning, cross-dataset, and domain generalization tasks, surpassing the existing state-of-the-art by a significant margin. Additionally, the study includes extensive ablation analysis to confirm the effectiveness of each proposed component and explore feasible alternatives. We believe that our consistency-guided tuning approach will have a substantial impact on tuning vision and vision-language models for various applications.

References

[1] 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] 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.

[3] 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.

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[5] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Maple: Multi-modal prompt learning. *arXiv preprint arXiv:2210.03117*, 2022.

[6] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021.

[7] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021.

[8] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133, 2022.

[9] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021.

[10] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

[11] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

[12] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

[13] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*, 2021.

[14] Beier Zhu, Yulei Niu, Yucheng Han, Yue Wu, and Hanwang Zhang. Prompt-aligned gradient for prompt tuning. *arXiv preprint arXiv:2205.14865*, 2022.

[15] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18082–18091, 2022.

[16] Hyojin Bahng, Ali Jahanian, Swami Sankaranarayanan, and Phillip Isola. Exploring visual prompts for adapting large-scale models. *arXiv preprint arXiv:2203.17274*, page 4, 2022.

[17] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*, 2022.

[18] Hantao Yao, Rui Zhang, and Changsheng Xu. Visual-language prompt tuning with knowledge-guided context optimization. *arXiv preprint arXiv:2303.13283*, 2023.