Generalizable Meta Prompt Learning for Vision-Language Model

Jeong-Jin Kim[1], Jung-Seul Ok[2]
[1]Master Student, Dept. of Computer Science, Pohang University of Science and Technology
[2]Professor, Dept. of Computer Science, Pohang University of Science and Technology

refstd@postech.ac.kr, jungseul.ok@postech.ac.kr

Abstract

Vision-language models (VLMs) are highly sensitive to prompt phrasing, which can substantially affect performance. While automatic prompt tuning methods improve performance, they still struggle to generalize to unseen datasets, particularly in low-data scenarios. We identify that suboptimal prompt initialization is a critical factor behind these generalization challenges. To address this, we propose a lightweight meta-prompt learning framework that uses Model-Agnostic Meta-Learning (MAML) combined with gradient clipping to train transferable prompts under simulated few-shot settings. Unlike prior approaches that require large-scale pretraining and introduce auxiliary networks, our method preserves the original prompt tuning structure, offering better interpretability and practical applicability. Through experiments on 11 diverse datasets, we show that our meta-prompts consistently outperform handcrafted prompts in zero-shot classification and, when used for fine-tuning, maintain or even improve generalization performance. Our findings highlight the potential of simple, data-efficient meta-prompt strategies to improve robustness and transferability in prompt tuning.

1. Introduction

Vision-language models (VLMs) rely heavily on textual prompts to guide image classification. Even minor variations—for example, changing “a photo of [CLASS]” to “a photo of a [CLASS]”—can lead to significant differences in performance. This makes finding effective prompts critical [1]. Traditionally, such prompts have been handcrafted through extensive trial and error. To alleviate this burden, gradient-based prompt tuning methods like CoOp have emerged, offering significant performance gains by learning prompts automatically [1]. Successive models like CoCoOp [2] and PromptSRC [3] further improved generalization via regularization strategies.

Despite these advances, prompt tuning still struggles to generalize to unseen datasets, especially under low-data conditions. While some methods have attempted to reduce overfitting, a major factor behind poor generalization is suboptimal prompt initialization. This underscores the need for a more robust and transferable initialization strategy—what we refer to as a “meta-prompt”—that provides a strong starting point across diverse tasks.

Recent methods such as GRAM [6] have explored this direction by combining meta-learning with gradient regulation to improve prompt generalization. However, these approaches typically rely on large-scale image-text pretraining and require modifications to existing prompt tuning methods by introducing an additional gradient-regulating network into the optimization process. In contrast, we propose a lightweight alternative that simulates low-data scenarios using a small set of labeled datasets. Our method preserves the original prompt tuning framework, offering better interpretability, simpler implementation, and improved applicability to real-world scenarios.

2. Related Work

2.1. Prompt Tuning in Vision-Language Models

Prompt tuning has emerged as an effective approach to adapt pre-trained vision-language models (VLMs) such as CLIP [5] to various downstream tasks without requiring full model fine-tuning. In image classification tasks, CLIP classifies images based on the cosine similarity ranking between vision and text features which are generated from the input image, class labels, and prompts—enabling the classification of unseen data, i.e., zero-shot classification [1,5].

CoOp introduced learnable continuous prompts optimized via gradient descent, demonstrating significant improvements over handcrafted prompts for such tasks [1]. However, CoOp exhibited limited generalization to unseen data, as the learned prompts tended to overfit to the training distribution.

To address this limitation, CoCoOp proposed a conditional prompt learning strategy where learnable prompts are aggregated with input-image dependent features generated by a simple neural network [2]. While effective, CoCoOp still relied on shallow conditioning mechanisms, which could be insufficient for highly complex domains.

Building upon these foundations, MaPLe introduced a multi-modal prompt learning paradigm that simultaneously tunes both vision and language branches of VLMs. [8]. Further enhancing generalization, PromptSRC introduced source-aware regularization built upon a variation of this visual-text prompt learning framework [3]. By regularizing the prompts to stay aligned with the original classification features, PromptSRC reduces overfitting and improves overall performance.

2.2 Meta-Learning for Prompt Initialization

Meta-learning offers a solution by enabling models to learn how to adapt prompts across diverse tasks. Meta-learning, or "learning to learn," emerged to address the limitations of traditional machine learning models that often require extensive data and training to perform well on new tasks. By leveraging experience from a variety of tasks, meta-learning aims to enable models to adapt quickly to new, unseen tasks

with minimal data, mimicking the human ability to learn efficiently from limited examples [4].

To discover a generalizable prompt initialization, GRAM introduced a gradient regulated meta-prompt learning framework to mitigate overfitting and improve cross-domain generalization [6]. They train both the meta-prompt and the weights of the gradient-regulating network by simulating how the trained prompts generalize to unseen data. Afterward, the trained meta-prompt is used as the initialization for fine-tuning, and the frozen weights of the gradient-regulating network are applied at this stage.

2.3 Our Contributions

While GRAM benefits from large-scale pretraining [6], our approach is specifically designed for small-data settings. Moreover, GRAM introduces an auxiliary network to regulate gradients and modifies existing methods by integrating this network into the gradient application process [6]. In contrast, our method adopts a simpler approach by applying a gradient clipping strategy during meta-prompt training, without altering the underlying method.

Compared to GRAM, our main contributions are as follows:

- We propose a lightweight meta-prompt training framework that simulates low-data environments without relying on large-scale pretraining.
- We introduce a gradient clipping strategy instead of adding new networks, maintaining full compatibility with existing prompt tuning methods.
- Our method enables clearer analysis of prompt initialization effects by preserving the original architecture.

3. Method

3.1. MAML for few-shot meta-prompt learning

We employ MAML [4] to train a generalizable prompt initialization, referred to as the meta-prompt. The training process is illustrated in Algorithm 1.

Suppose we have a set of different N-shot learning tasks $\mathcal{T}_i \in \mathcal{T}^{tr}$ that contains training data. First, we simulate task specific learning initialized with the current meta-prompt p. For each task $\mathcal{T}_i$, we draw a small support set $\mathcal{D}_i$ of labeled datapoints from the N-shot training data. After the task-specific loss $\mathcal{L}_{\mathcal{T}_i}\big(f(\mathcal{D}_i; p, \theta)\big)$ is computed, a prompt adaptation step is performed using gradient descent to obtain a task-specific prompt p_i'.

To assess and enhance generalization capability of the fine-tuned prompt p_i' which is initialized from the current meta-prompt p, we compute the generalization loss $\mathcal{L}_{\mathcal{T}_i}\big(f(\mathcal{D}_i'; p_i', \theta)\big)$ on a query set $\mathcal{D}_i'$ which is disjoint to $\mathcal{D}_i$. This loss measures how well the fine-tuned prompt p_i' generalizes to unseen data beyond the fine-tuning set. By averaging this loss across all tasks, we can calculate the main gradient $\mathcal{G}$ with respect to p for each task and update p using gradient descent. We apply the Adam optimizer [7] along with gradient clipping into $\mathcal{G}$ to ensure stable updates.

3.2. Why Gradient Clipping is Necessary

Gradient clipping is essential for stable training in our low-data meta-prompt setting. Without it, we observed that excessively large gradients cause abrupt parameter updates that push the model into suboptimal, dead-end regions of the parameter space. Once in these regions, the model struggles to recover, resulting in stalled learning. This failure not only hampers the initial training but also propagates to downstream tasks, leading to ineffective or failed fine-tuning. Gradient clipping mitigates this risk by constraining the magnitude of updates, ensuring smoother optimization and effective fine-tuning.

Algorithm 1 MAML for Few-Shot Meta-Prompt Learning

Require: $\mathcal{T}^{tr}$: N-shot learning tasks
Require: f: VLM like CLIP
Require: $\mathcal{L}$: soft prompt loss function
Require: α, β: step size hyperparameters
Require: γ: gradient clipping threshold hyperparameter
Require: Frozen parameters θ for f

Initialize prompt p with a handcrafted prompt
while not converges **do**
  for all $\mathcal{T}_i \in \mathcal{T}^{tr}$ **do**
    Randomly divide $\mathcal{T}_i$ into K datapoints $\mathcal{D}_i$ and $N-K$ datapoints $\mathcal{D}_i'$
    Evaluate $\nabla_p \mathcal{L}_{\mathcal{T}_i}(f(D_i; p, \theta))$ using $\mathcal{D}_i$ and $\mathcal{L}_{\mathcal{T}_i}$
    Compute the adapted prompt with gradient descent:
    $p_i' = p - \alpha \nabla_\theta \mathcal{L}_{\mathcal{T}_i}(f(\mathcal{D}_i; p, \theta))$
  end for
  Compute gradient $\mathcal{G} = \nabla_p \sum_{\mathcal{T}_i \in \mathcal{T}^{tr}} \mathcal{L}_{\mathcal{T}_i}(f(\mathcal{D}_i'; p_i', \theta))$ using each $\mathcal{D}_i'$ and $\mathcal{L}_{\mathcal{T}_i}$
  Update $p \leftarrow p - \mathrm{Adam}(\beta)\frac{\gamma}{\max\{||\mathcal{G}||_2, \gamma\}}\mathcal{G}$
end while

4. Experiments

4.1. Experimental Setup

We evaluate our approach on 11 publicly available image classification datasets, as shown in Table 1. The datasets cover a diverse range of tasks such as object, scene, and action recognition. These datasets are manually split based on known characteristics to ensure balanced feature distribution between training and test sets. Also, we focus on small-scale datasets to enable fast training and better highlight generalization capability.

Table 1: Dataset Segmentation for MetaPrompt Training

Purpose	Datasets			
	Train	Classes	Test	Classes
General recognition Object recognition	Caltech101	100	ImageNet	1000
Scene recognition Action recognition	UCF101	101	SUN397	397
Specialized tasks	EuroSAT	10	DTD	47
Fine-grained datasets	FGVC Aircraft Oxford Pets Flowers102	100 37 102	Stanford Cars Food101	196 101
Total		450		1741

To assess task-specific generalization performance, we fine-tune meta-prompts using the same method as in the initial training stage. For this evaluation, each dataset is split into base and novel classes based on their labels, with novel classes excluded from the fine-tuning process. The number of training images per class is kept the same as in the previous stage.

We compare the performance of two fine-tuning setups: one initialized with meta-prompts obtained from the initial training stage, and the other initialized with a handcrafted prompt (“a photo of a”). All other settings, including the fine-

tuning procedure and data splits, are kept identical to ensure a fair comparison of initialization strategies.

4.2. Implementational Details

We implement three variants of our meta-prompt—MP-CoOp, MP-CoCoOp, and MP-PromptSRC—by combining MAML with the respective tuning method (CoOp, CoCoOp, PromptSRC) on ViT/B16 CLIP [5]. Each task $\mathcal{T}_i$ corresponds to a training dataset and is constructed with N = 16 and K = 5 (Total 449 classes, 16 images for each). Every learnable parameter of the existing methods is trained and subsequently used for initialization. These meta-prompts can either be used as-is, following the way the original method performs inference, or be used to initialize the corresponding fine-tuning method.

For meta-prompt learning, each meta-prompt is initialized with "a photo of a" and trained for 200 epochs with $\alpha = 0.025/6$, $\beta = 0.001$. $\gamma = 5$ ($\alpha_{MP-PromptSRC} = 0.02/6$). The inner update to compute p_i' is repeated 4 times. Batch sizes of the inner update are 32, 8, 32 for MP-CoOp, MP-CoCoOp, MP-PromptSRC, respectively. For fine-tuning, the number of epochs is 200, 10, 20, learning rates are 0.002, 0.002, 0.0025, batch sizes are 32, 1, 4 for CoOp, CoCoOp, PromptSRC, respectively. Common settings are as follows: for CoOp and CoCoOp, prompt length is 4, class token position is end, and class specific context (CSC) is disabled. For PromptSRC, length of vision and text prompt is 4, their depth is 9. GPA mean is 15 and standard variation is 1. Text loss weight is 25 and image loss weight is 20.

4.3. Results

Table 2 summarizes the experimental results. Our proposed meta-prompts—MP-CoOp, MP-CoCoOp, and MP-PromptSRC—are evaluated against both handcrafted prompts and their corresponding base prompt-tuning methods (CoOp, CoCoOp, PromptSRC). The results demonstrate that all meta-prompts outperform handcrafted prompts when used directly. However, with fine-tuning, while MP-CoOp failed to generalize to unseen data, the others showed comparable performance to handcrafted prompt initialization.

4.3.1 Zero-shot Performance

In the zero-shot setting, all meta-prompts consistently outperform handcrafted prompts. Among them, MP-PromptSRC yields the most notable gains, achieving an improvement of +4.83%p (70.09%) across all classes, compared to 65.26% with the handcrafted prompt. On the test set, MP-PromptSRC likewise outperforms its handcrafted counterpart, reaching +1.04%p (66.10%) versus 65.06%. However, the performance gain on the test set is less pronounced than on the training set, and MP-CoCoOp shows a slight performance decline (-0.28%p) on the test set.

4.3.2 Fine-tuning Performance

When fine-tuning is applied, meta-prompts exhibit varying performance. Although MP-CoOp improves on the base metric, it suffers a significant performance drop — over 10 %p — on both the novel and all metrics. These results align with Zhou [2], who demonstrated that CoOp is vulnerable to overfitting.

In contrast, MP-CoCoOp and MP-PromptSRC experience slight decreases in the base metric, by -0.44%p and -2.09%p respectively, while achieving performance gains on the novel metric, with increases of +0.42%p and +0.44%p. For the all metric, only MP-CoCoOp shows an improvement, with a gain of +0.62%p.

5. Discussion & Conclusion

We demonstrated that meta-prompts trained with a small amount of data can outperform handcrafted prompts and achieve improved generalization performance. Furthermore, when using meta-prompts as initialization for fine-tuning, CoOp exhibited a decline in generalization performance, whereas CoCoOp and PromptSRC maintained performance levels comparable to their original results. Although fine-tuning the meta-prompts did not lead to substantial improvements, likely due to the limited amount of training data, our results highlight the strong potential of meta-prompts even in low-data regimes. These findings reaffirm the importance of efforts to prevent overfitting in prompt tuning and highlight meta-prompts as a robust and practical alternative to handcrafted prompts.

Acknowledgement

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2019-II191906, Artificial Intelligence Graduate School Program (POSTECH))

References

[1] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," International Journal of Computer Vision, vol. 130, no. 9, pp. 2337–2348, 2022.

[2] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 16816–16825, June 2022.

[3] M. U. Khattak, S. T. Wasim, M. Naseer, S. Khan, M.-H. Yang, and F. S. Khan, "Self-regulating prompts: Foundational model adaptation without forgetting," in Proceedings of the IEEE/CVF international conference on computer vision, pp. 15190–15200.

[4] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta- learning for fast adaptation of deep networks," in Proceedings of the 34th International Conference on Machine Learning (D. Precup and Y. W. Teh, eds.), vol. 70 of Proceedings of Machine Learning Research, pp. 1126–1135, PMLR, 06–11 Aug 2017.

[5] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in Proceedings of the 38th International Conference on Machine Learning (M. Meila and T. Zhang, eds.), vol. 139 of Proceedings of Machine Learning Research, pp. 8748–8763, PMLR, 18–24 Jul 2021.

[6] J. Li, M. Gao, L. Wei, S. Tang, W. Zhang, M. Li, W. Ji, Q. Tian, T.-S. Chua, and Y. Zhuang, "Gradient-regulated meta-prompt learning for generalizable vision-language models," in Proceedings of the IEEE/CVF international conference on computer vision, pp. 2551–2562, 2023.

[7] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[8] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. "Maple: Multi-modal prompt learning" in CVPR, pp. 19113– 19122, 2023.

[9] Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. "Fine-tuned clip models are efficient video learners." in CVPR, pp. 6545–6554, 2023.

Table 2: **Comparison of Base-to-Novel Generalization Performance Gains Using Meta-Prompts.** H denotes the performance using handcrafted prompts, shown as absolute values. MP denotes meta-prompts, with values presented as relative performance changes (e.g., -2.09 %p, +0.44 %p) compared to H. The datasets used for meta-prompt training are in **bold**. In the zero-shot setting, meta-prompts (CoOp, CoCoOp, PromptSRC) are evaluated against the handcrafted baseline. In the fine-tuning setting, each meta-prompt is compared to its corresponding handcrafted prompt initialization, using the same training approach.

Fine-Tuning		Zero-Shot				Fine-Tune					
Method		-	CoOp	CoCoOp	PromptSRC	CoOp		CoCoOp		PromptSRC	
Initialization		H	MP	MP	MP	H	MP	H	MP	H	MP
Dataset	Metric	Accuracy (%)									
Average	Base	69.35	+2.85	+1.89	+5.15	78.91	+0.55	78.63	-0.44	84.02	-2.09
	New	74.24	+0.03	+1.27	+2.81	83.01	-18.71	72.22	+0.42	75.39	+0.44
	All	65.26	+1.85	+1.75	+4.83	75.58	-11.52	68.32	+0.62	73.41	-0.54
Average (Train)	Base	69.07	+4.13	+3.53	+8.43	77.97	-0.02	78.59	-1.21	85.28	-4.17
	New	74.38	+0.60	+3.57	+4.80	83.66	-18.38	71.38	+0.94	75.05	+1.48
	All	65.43	+3.07	+3.45	+7.98	75.48	-12.37	67.66	+0.99	74.35	-0.92
Average (Test)	Base	69.70	+1.30	-0.08	+1.22	80.04	+1.24	78.68	+0.48	82.50	+0.40
	New	74.06	-0.66	-1.48	+0.42	82.22	-19.10	73.23	-0.21	75.80	-0.82
	All	65.06	+0.38	-0.28	+1.04	75.71	-10.51	69.11	+0.19	72.28	-0.08
Caltech101	Base	97.00	-0.70	-0.80	+0.00	97.77	+0.43	97.80	-0.40	97.90	+0.40
	New	94.00	+1.20	+1.40	+1.10	95.77	-6.57	92.90	+2.20	93.60	+1.20
	All	92.90	+1.30	+0.40	+1.40	95.50	-4.10	93.73	+0.67	94.60	+0.10
EuroSAT	Base	56.40	+10.60	+15.00	+36.10	92.87	-0.17	85.83	-0.33	91.60	+2.80
	New	63.80	-2.20	+19.20	+15.50	92.10	-35.90	61.33	-1.33	69.40	+10.20
	All	47.70	+6.20	+15.20	+30.90	83.70	-30.10	56.40	+4.30	70.30	+6.20
FGVC Aircraft	Base	27.10	+4.30	+3.20	+5.30	38.03	+1.87	35.13	-0.83	42.60	+0.00
	New	36.20	+3.00	-0.30	+9.20	57.00	-30.80	32.17	-1.07	36.80	-1.70
	All	24.80	+3.10	+0.60	+6.40	40.40	-14.20	26.27	-0.87	31.90	-0.20
Flowers102	Base	72.20	+2.90	-1.80	-1.70	60.40	+0.00	75.57	-6.37	97.90	-29.00
	New	77.90	-0.40	-0.10	-1.20	71.90	-10.60	71.47	+2.83	77.00	-2.80
	All	71.40	+0.00	-1.00	-1.60	59.87	-7.17	66.50	+0.00	80.80	-13.60
OxfordPets	Base	91.20	+3.50	+3.10	+4.40	94.33	-2.03	95.13	+0.77	95.50	+0.20
	New	97.00	+1.00	+0.80	+0.80	97.33	-1.73	97.93	-0.43	97.50	-0.20
	All	89.10	+3.90	+2.70	+4.00	92.40	-5.20	92.03	+0.47	92.70	+0.10
UCF101	Base	70.50	+4.20	+2.50	+6.50	84.40	-0.20	82.10	-0.10	86.20	+0.60
	New	77.40	+1.00	+0.40	+3.40	87.87	-24.67	72.47	+3.43	76.00	+2.20
	All	66.70	+3.90	+2.80	+6.80	81.03	-13.43	71.03	+1.37	75.80	+1.90
DTD	Base	53.20	+3.30	-2.00	+3.70	78.30	+4.50	77.10	+1.10	82.80	+2.30
	New	60.40	-3.50	-4.40	-1.90	75.83	-31.03	53.53	-0.83	62.90	-4.90
	All	44.50	-1.10	-0.90	+0.90	67.93	-17.23	51.57	+0.23	58.70	-0.80
Food101	Base	90.10	-0.70	-0.40	-0.30	88.73	-0.53	90.63	-0.33	90.90	-0.30
	New	91.30	-0.20	-0.40	+0.20	90.80	-7.30	91.33	-0.33	91.70	-0.30
	All	86.10	-0.50	-0.70	-0.20	85.00	-4.50	86.40	-0.40	86.70	-0.30
ImageNet	Base	72.40	+2.10	+1.40	+1.30	75.80	+0.50	75.90	-0.30	77.70	-0.20
	New	68.10	+1.60	+1.20	+2.00	73.63	-10.63	70.57	-0.37	70.60	+0.00
	All	66.70	+1.80	+1.30	+1.70	71.50	-5.30	69.87	-0.37	70.70	-0.10
Stanford Cars	Base	63.50	+0.20	-0.90	-0.30	76.77	+1.63	70.40	+1.50	78.50	+0.20
	New	75.00	-0.50	-1.10	+1.20	87.23	-25.03	73.63	-0.33	75.20	+1.00
	All	65.40	+0.30	-1.40	+0.60	79.57	-13.17	68.27	+0.53	73.40	+0.50
SUN397	Base	69.30	+1.60	+1.50	+1.70	80.60	+0.10	79.37	+0.43	82.60	+0.00
	New	75.50	-0.70	-2.70	+0.60	83.60	-21.50	77.07	+0.83	78.60	+0.10
	All	62.60	+1.40	+0.30	+2.20	74.53	-12.33	69.47	+0.93	71.90	+0.30