DMQA Open Seminar

Prompt Learning in Vision-Language Models: From Text Prompts to Multimodal Representations

2026. 07. 24.

Data Mining & Quality Analytics Lab.

손병우

About Me

Researcher Profile

❖ 손병우 (Son Byeong Woo)

- 고려대학교 산업경영공학과 석사과정 (2026.03~Present)
- Data Mining & Quality Analytics Lab. (김성범 교수님)

❖ 관심분야

- Vision-Language Model
- Representation Learning

❖ Contact

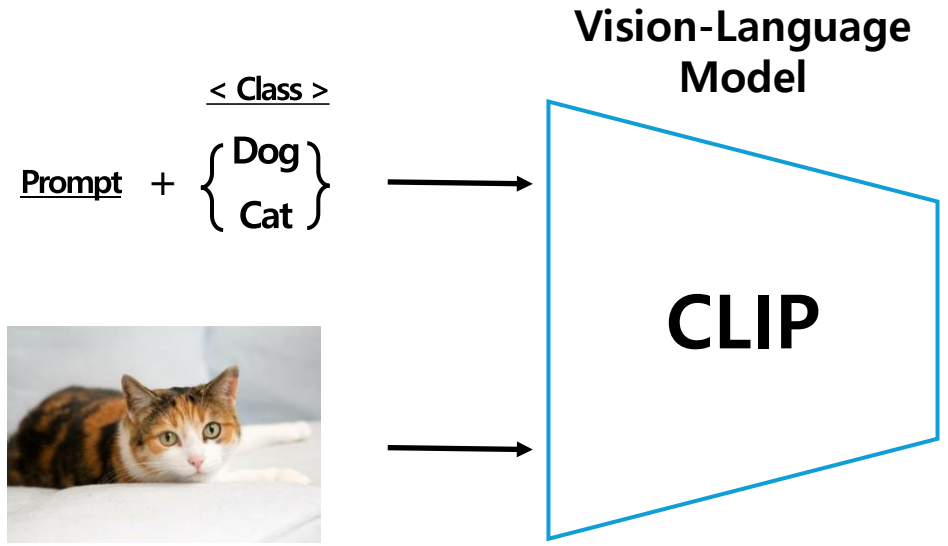
- rooneei@korea.ac.kr

Introduction

Backgroud

❖ Prompt Learning 이란 무엇일까?

- **Prompt :** Vision-Language Model에서 이미지와의 유사도를 계산하기 위해 자연어 문장 형태로 표현한 텍스트 입력
- **Prompt Learning** : Prompt + Learning

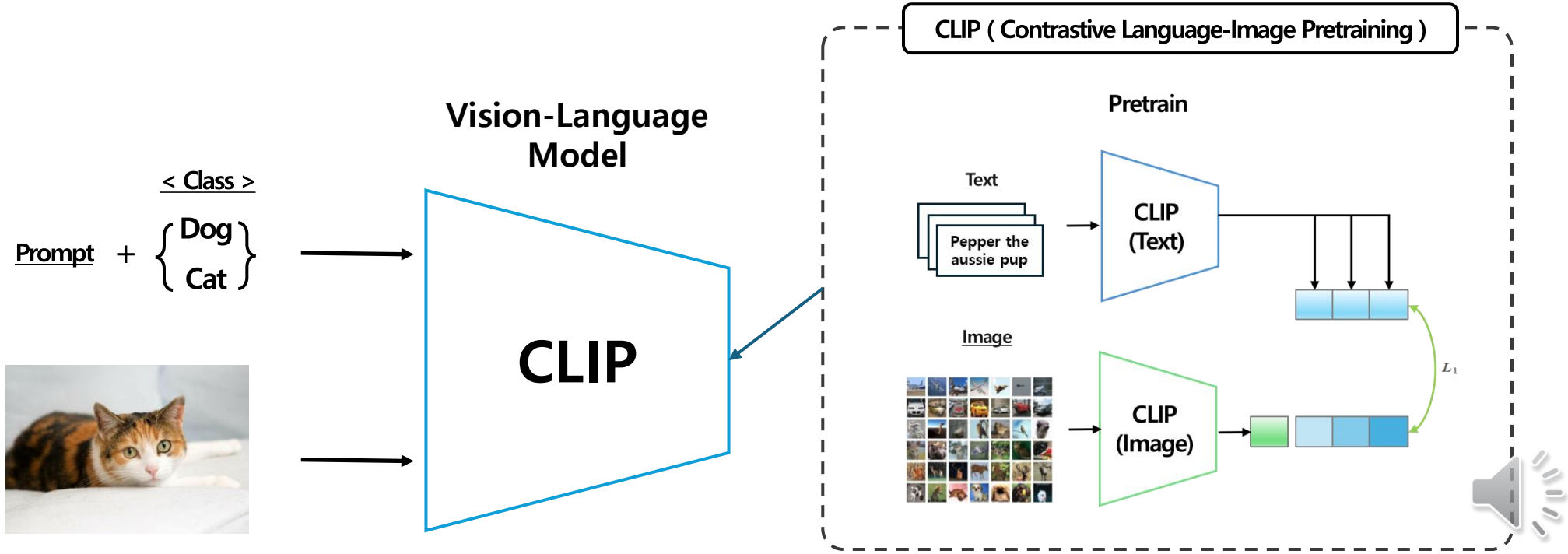


Introduction

Backgroud

❖ Prompt Learning 이 왜 필요할까?

- **Prompt :** Vision-Language Model에서 이미지와의 유사도를 계산하기 위해 자연어 문장 형태로 표현한 텍스트 입력
- **CLIP** (VLM, Vision-Language Model) **:** 대규모 이미지–텍스트 쌍으로 사전 학습된 대표적인 파운데이션 VLM 모델로, 이미지-텍스트 표현 간 유사도를 비교하여 **별도의 추가 학습 없이 Zero-shot 이미지 분류를 수행**할 수 있음

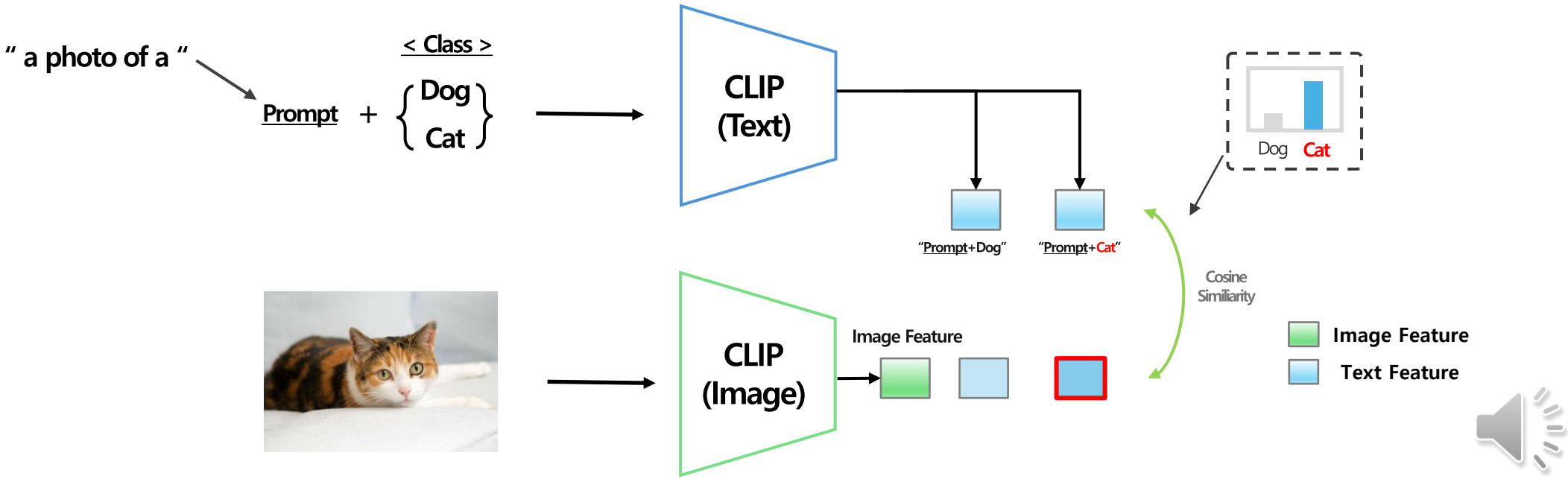


Introduction

Backgroud

❖ Prompt Learning 이 왜 필요할까?

- **Prompt :** Vision-Language Model에서 이미지와의 유사도를 계산하기 위해 자연어 문장 형태로 표현한 텍스트 입력
- **CLIP** (VLM, Vision-Language Model) **:** 대규모 이미지–텍스트 쌍으로 사전 학습된 대표적인 파운데이션 VLM 모델로, 이미지-텍스트 표현 간 유사도를 비교하여 **별도의 추가 학습 없이 Zero-shot 이미지 분류를 수행**할 수 있음



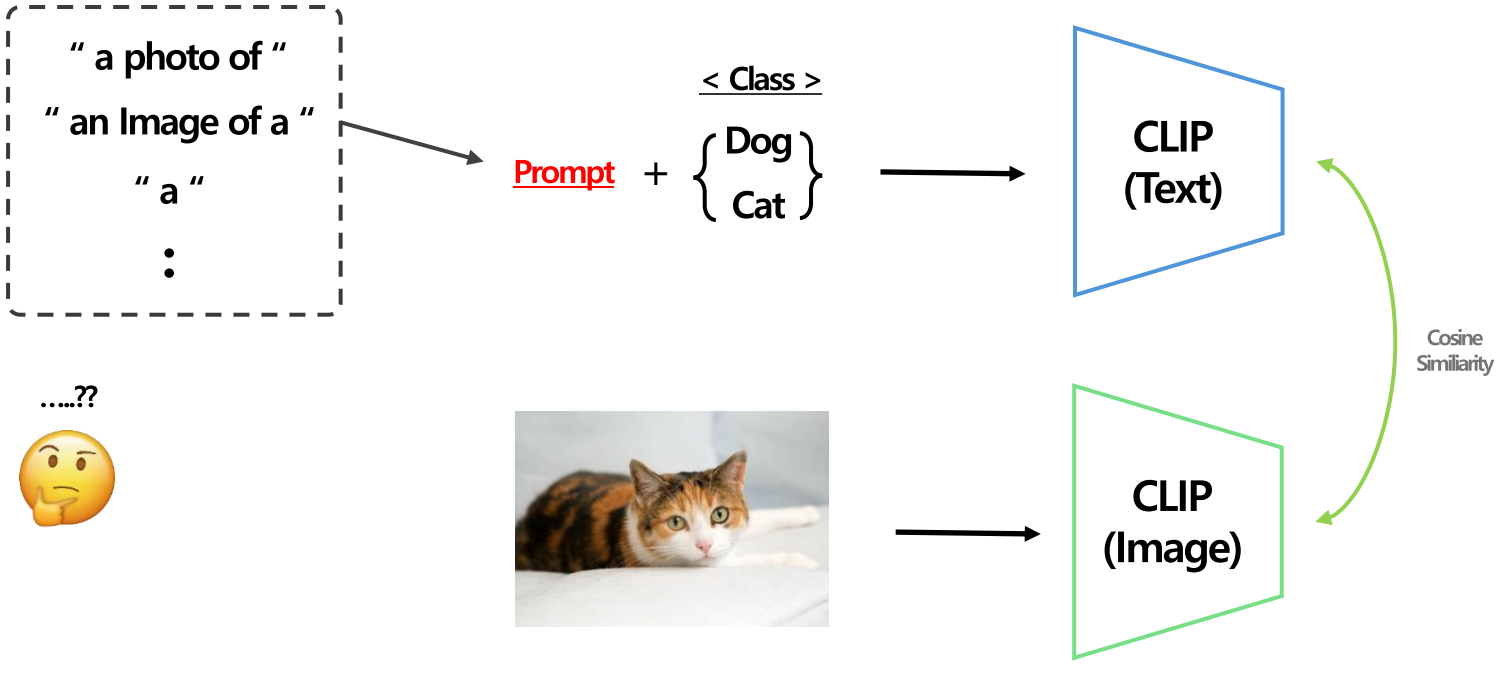
Introduction

Backgroud

❖ Prompt Learning 이 왜 필요할까?

- **문제 상황** : 동일한 Classification Task 에서도 **Prompt 에 따라 이미지–텍스트 정렬 및 분류 성능이 달라질 수 있음**

  → 데이터셋에 따른 최적의 Prompt를 찾기 위해 다양한 문장을 반복적으로 설계·평가하는데 많은 비용이 요구됨



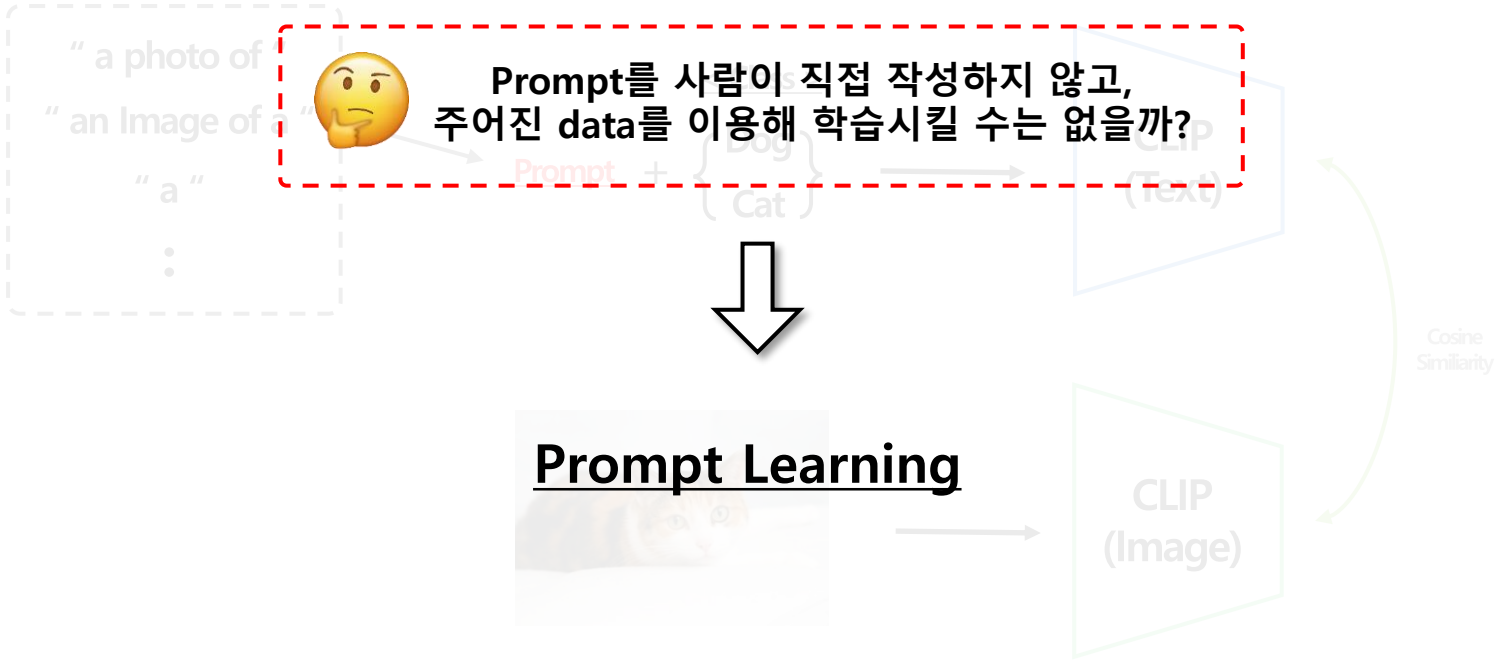
Introduction

Backgroud

❖ Prompt Learning 이 왜 필요할까?

- **문제 상황** : 동일한 Classification Task 에서도 **Prompt 에 따라 이미지–텍스트 정렬 및 분류 성능이 달라질 수 있음**

  → 데이터셋에 따른 최적의 Prompt를 찾기 위해 다양한 문장을 반복적으로 설계·평가하는데 많은 비용이 요구됨



**Prompt를 사람이 직접 작성하지 않고,
주어진 data를 이용해 학습시킬 수는 없을까?**

Prompt Learning

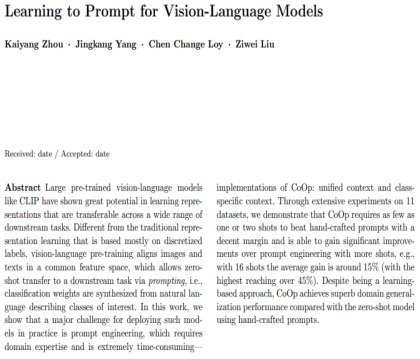
Introduction

Backgroud

❖ **Prompt Learning 은 어떠한 발전을 이뤄왔을까?**

Text Prompt | **Multimodal Prompt** | **Multimodal Representation**

CoOp [IJCV 2022] → **CoCoOp** [CVPR 2022] → **MaPLe** [CVPR 2023] → **MMRL** [CVPR 2025]



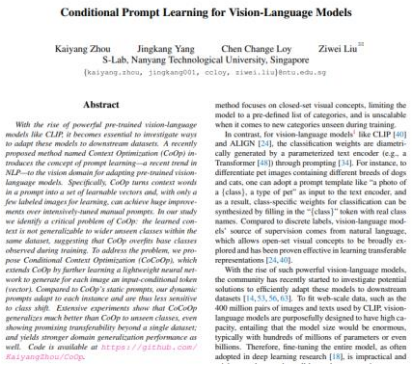
Learning to Prompt for Vision-Language Models

Kaiyang Zhou · Jingkang Yang · Chen Change Loy · Ziwei Liu

Received: date / Accepted: date

Abstract Large pre-trained vision-language models like CLIP have shown great potential in learning representations that are transferable across a wide range of downstream tasks. Different from the traditional representation learning that is based mostly on discretized labels, vision-language pre-training aligns images and texts in a common feature space, which allows zero-shot transfer to a downstream task via *prompting*, i.e., classification weights are synthesized from natural language describing classes of interest. In this work, we show that a major challenge for deploying such models in practice is prompt engineering, which requires domain expertise and is extremely time-consuming—

implementations of CoOp: unified context and class-specific context. Through extensive experiments on 11 datasets, we demonstrate that CoOp requires as few as one or two shots to beat hand-crafted prompts with a decent margin and is able to gain significant improvements over prompt engineering with more shots, e.g., with 16 shots the average gain is around 15% (with the highest reaching over 45%). Despite being a learning-based approach, CoOp achieves superb domain generalization performance compared with the zero-shot model using hand-crafted prompts.



Conditional Prompt Learning for Vision-Language Models

Kaiyang Zhou Jingkang Yang Chen Change Loy Ziwei Liu

S-Lab, Nanyang Technological University, Singapore

{kaiyang.zhou, jingkang001, ccloy, ziwei.liu}@ntu.edu.sg

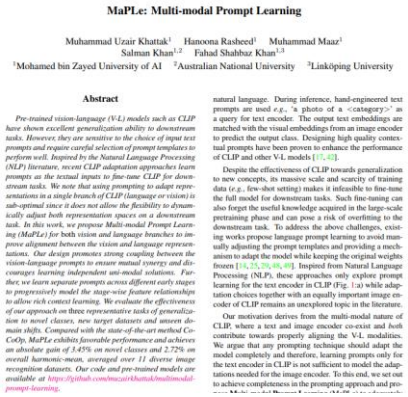
Abstract

With the rise of powerful pre-trained vision-language models like CLIP, it becomes essential to investigate ways to adapt these models to downstream datasets. A recently proposed method named Context Optimization (CoOp) introduces the concept of prompt learning—a recent trend in NLP—to the vision domain for adapting pre-trained vision-language models. Specifically, CoOp turns context words in a prompt into a set of learnable vectors and, with only a few labeled images for learning, can achieve huge improvements over intensively-tuned manual prompts. In our study we identify a critical problem of CoOp: the learned context is not generalizable to wider unseen classes within the same dataset, suggesting that CoOp overfits base classes observed during training. To address the problem, we propose Conditional Context Optimization (CoCoOp), which extends CoOp by further learning a lightweight neural network to generate for each image an input-conditional token (vector). Compared to CoOp's static prompts, our dynamic prompts adapt to each instance and are thus less sensitive to class shift. Extensive experiments show that CoCoOp generalizes much better than CoOp to unseen classes, even showing promising transferability beyond a single dataset; and yields stronger domain generalization performance as well. Code is available at https://github.com/KaiyangZhou/CoOp.

method focuses on closed-set visual concepts, limiting the model to a pre-defined list of categories, and is unscalable when it comes to new categories unseen during training.

In contrast, for vision-language models[1] like CLIP [40] and ALIGN [24], the classification weights are diametrically generated by a parameterized text encoder (e.g., a Transformer [48]) through prompting [34]. For instance, to differentiate pet images containing different breeds of dogs and cats, one can adopt a prompt template like "a photo of a {class}, a type of pet" as input to the text encoder, and as a result, class-specific weights for classification can be synthesized by filling in the "{class}" token with real class names. Compared to discrete labels, vision-language models' source of supervision comes from natural language, which allows open-set visual concepts to be broadly explored and has been proven effective in learning transferable representations [24, 40].

With the rise of such powerful vision-language models, the community has recently started to investigate potential solutions to efficiently adapt these models to downstream datasets [14, 53, 56, 63]. To fit web-scale data, such as the 400 million pairs of images and texts used by CLIP, vision-language models are purposefully designed to have high capacity, entailing that the model size would be enormous, typically with hundreds of millions of parameters or even billions. Therefore, fine-tuning the entire model, as often adopted in deep learning research [18], is impractical and



MaPLe: Multi-modal Prompt Learning

Muhammad Uzair Khattak[1] Hanoona Rasheed[1] Muhammad Maaz[1]

Salman Khan[1,2] Fahad Shahbaz Khan[1,3]

[1]Mohamed bin Zayed University of AI [2]Australian National University [3]Linköping University

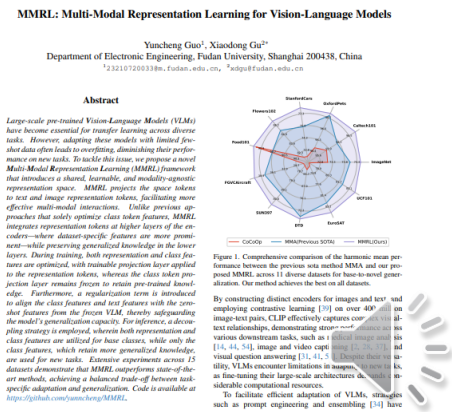
Abstract

Pre-trained vision-language (V-L) models such as CLIP have shown excellent generalization ability to downstream tasks. However, they are sensitive to the choice of input text prompts and require careful selection of prompt templates to perform well. Inspired by the Natural Language Processing (NLP) literature, recent CLIP adaptation approaches learn prompts as the textual inputs to fine-tune CLIP for downstream tasks. We note that using prompting to adapt representations in a single branch of CLIP (language or vision) is sub-optimal since it does not allow the flexibility to dynamically adjust both representation spaces on a downstream task. In this work, we propose Multi-modal Prompt Learning (MaPLe) for both vision and language branches to improve alignment between the vision and language representations. *Our design promotes strong coupling between the vision-language prompts to ensure mutual synergy and discourages learning independent uni-modal solutions. Further, we learn separate prompts across different early stages to progressively model the stage-wise feature relationships to allow rich context learning. We evaluate the effectiveness of our approach on* three *representative tasks of generalization to novel classes, new target datasets and unseen domain shifts. Compared with the state-of-the-art method CoCoOp, MaPLe exhibits favorable performance and achieves an absolute gain of 3.45% on novel classes and 2.72% on overall harmonic-mean, averaged over 11 diverse image recognition datasets. Our code and pre-trained models are available at https://github.com/muzairkhattak/multimodal-prompt-learning.*

natural language. During inference, hand-engineered text prompts are used *e.g.*, 'a photo of a <category>' as a query for text encoder. The output text embeddings are matched with the visual embeddings from an image encoder to predict the output class. Designing high quality contextual prompts have been proven to enhance the performance of CLIP and other V-L models [17, 42].

Despite the effectiveness of CLIP towards generalization to new concepts, its massive scale and scarcity of training data (e.g., few-shot setting) makes it infeasible to fine-tune the full model for downstream tasks. Such fine-tuning can also forget the useful knowledge acquired in the large-scale pretraining phase and can pose a risk of overfitting to the downstream task. To address the above challenges, existing works propose language prompt learning to avoid manually adjusting the prompt templates and providing a mechanism to adapt the model while keeping the original weights frozen [14, 25, 29, 48, 49]. Inspired from Natural Language Processing (NLP), these approaches only explore prompt learning for the text encoder in CLIP (Fig. 1a) while adaptation choices together with an equally important image encoder of CLIP remains an unexplored topic in the literature.

Our motivation derives from the multi-modal nature of CLIP, where a text and image encoder co-exist and *both* contribute towards properly aligning the V-L modalities. We argue that any prompting technique should adapt the model completely and therefore, learning prompts only for the text encoder in CLIP is not sufficient to model the adaptations needed for the image encoder. To this end, we set out to achieve completeness in the prompting approach and pro-



MMRL: Multi-Modal Representation Learning for Vision-Language Models

Yuncheng Guo[1], Xiaodong Gu[2*]

Department of Electronic Engineering, Fudan University, Shanghai 200438, China

[1]23210720033@m.fudan.edu.cn, [2]xdgu@fudan.edu.cn

Abstract

Large-scale pre-trained Vision-Language Models (VLMs) have become essential for transfer learning across diverse tasks. However, adapting these models with limited few-shot data often leads to overfitting, diminishing their performance on new tasks. To tackle this issue, we propose a novel Multi-Modal Representation Learning (MMRL) framework that introduces a shared, learnable, and modality-agnostic representation space. MMRL projects the space tokens to text and image representation tokens, facilitating more effective multi-modal interactions. Unlike previous approaches that solely optimize class token features, MMRL integrates representation tokens at higher layers of the encoders—where dataset-specific features are more prominent—while preserving generalized knowledge in the lower layers. During training, both representation and class features are optimized, with trainable projection layer applied to the representation tokens, whereas the class token projection layer remains frozen to retain pre-trained knowledge. Furthermore, a regularization term is introduced to align the class features and text features with the zero-shot features from the frozen VLM, thereby safeguarding the model's generalization capacity. For inference, a decoupling strategy is employed, wherein both representation and class features are utilized for base classes, while only the class features, which retain more generalized knowledge, are used for new tasks. Extensive experiments across 15 datasets demonstrate that MMRL outperforms state-of-the-art methods, achieving a balanced trade-off between task-specific adaptation and generalization. Code is available at https://github.com/yunncheng/MMRL.

Figure 1. Comprehensive comparison of the harmonic mean performance between the previous sota method MMA and our proposed MMRL across 11 diverse datasets for base-to-novel generalization. Our method achieves the best on all datasets.

By constructing distinct encoders for images and text and employing contrastive learning [39] on over 400 million on image-text pairs, CLIP effectively captures cor ful visual text relationships, demonstrating strong performance across various downstream tasks, such as zero-shot image analysis [14, 44, 54], image and video captioning [2, 28, 51], and visual question answering [31, 41, 5]. Despite their versatility, VLMs encounter limitations in adapting to new tasks, as fine-tuning their large-scale architectures demands considerable computational resources.

To facilitate efficient adaptation of VLMs, strategies such as prompt engineering and ensembling [34] have

1. CoOp

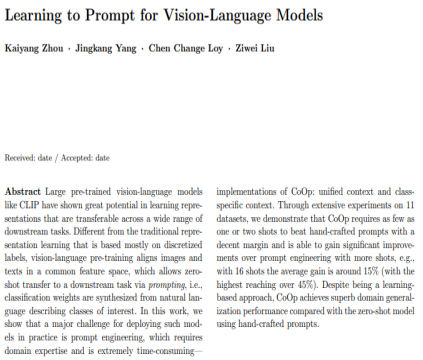
Context Optimization (CoOp)

❖ **Prompt Learning 은 어떠한 발전을 이뤄왔을까?**

CoOp

[IJCV 2022]

Text Prompt Tuning



Learning to Prompt for Vision-Language Models

Kaiyang Zhou · Jingkang Yang · Chen Change Loy · Ziwei Liu

Received: date / Accepted: date

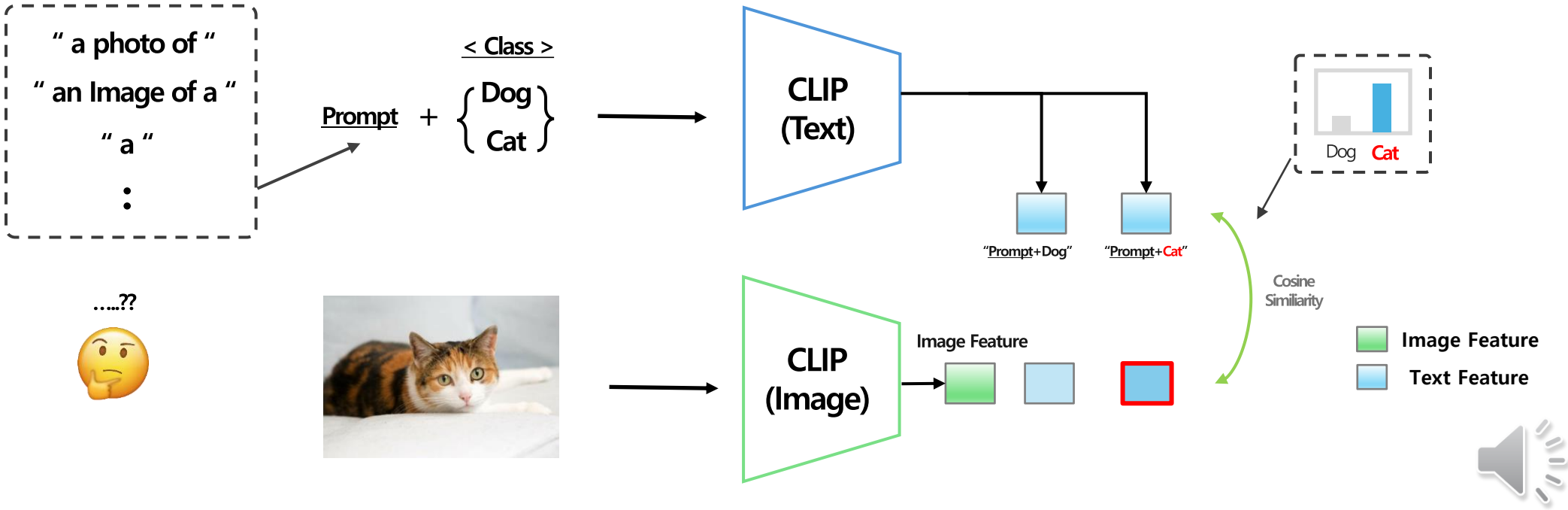
Abstract Large pre-trained vision-language models like CLIP have shown great potential in learning representations that are transferable across a wide range of downstream tasks. Different from the traditional representation learning that is based mostly on discretized labels, vision-language pre-training aligns images and texts in a common feature space, which allows zero-shot transfer to a downstream task via *prompting*, i.e., classification weights are synthesized from natural language describing classes of interest. In this work, we show that a major challenge for deploying such models in practice is prompt engineering, which requires domain expertise and is extremely time-consuming—implementations of CoOp: unified context and class-specific context. Through extensive experiments on 11 datasets, we demonstrate that CoOp requires as few as one or two shots to beat hand-crafted prompts with a decent margin and is able to gain significant improvements over prompt engineering with more shots, e.g., with 16 shots the average gain is around 15% (with the highest reaching over 45%). Despite being a learning-based approach, CoOp achieves superb domain generalization performance compared with the zero-shot model using hand-crafted prompts.

1. CoOp

Context Optimization (CoOp)

❖ Learning to Prompt for Vision-Language Model [IJCV 2022]

- **문제 상황** : 데이터셋에 따른 최적의 Prompt를 찾기 위해 다양한 문장을 반복적으로 설계·평가하는데 많은 비용이 요구됨
- " 사람이 직접 만든 Prompt 처럼, 주어진 몇 장의 이미지만으로 자동화된 Prompt 를 학습시킬 수는 없을까? "

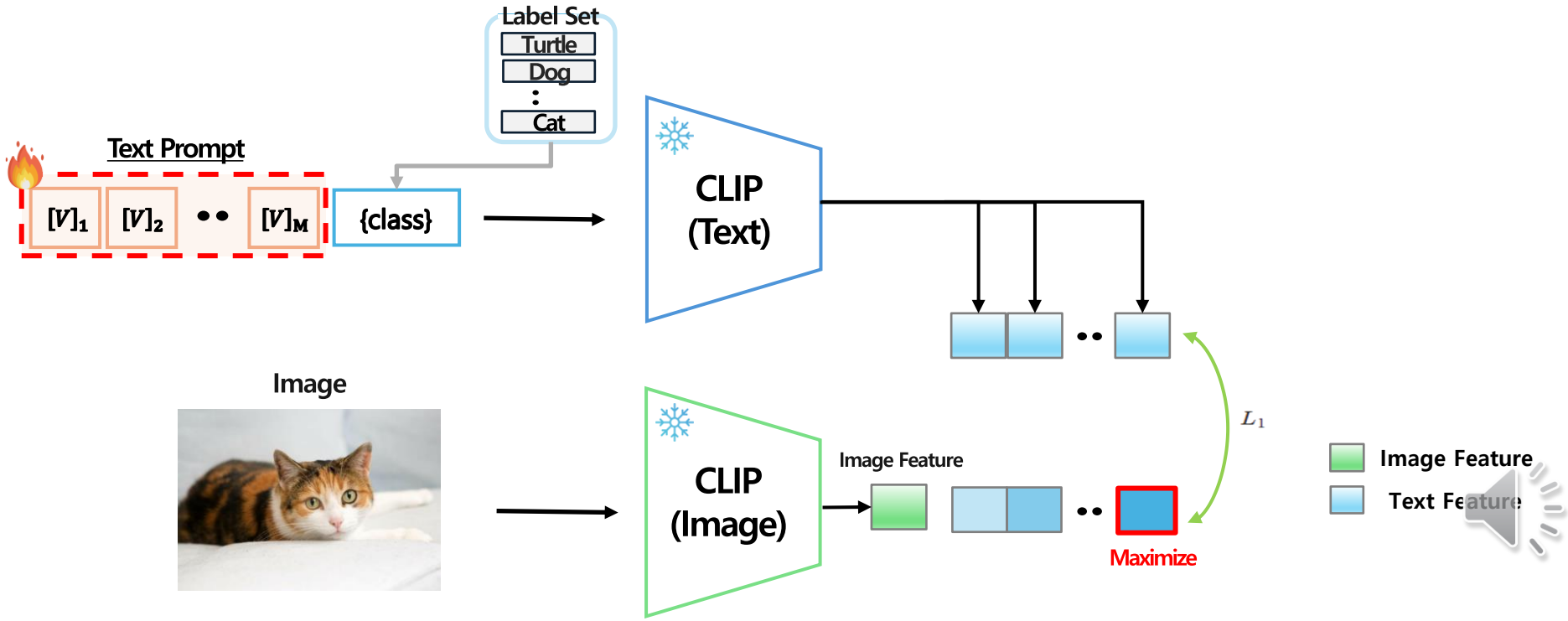


1. CoOp

Context Optimization (CoOp)

❖ Method & Result

- **주요 기여** : 기존 CLIP의 handcrafted prompt를 학습 가능한 continuous context vector $[v_1, \ldots, v_M]$ 로 대체
- Parameter-Efficient Learning : CLIP encoder는 frozen 상태로 유지하고 context vector만 학습 → 적은 파라미터 수로 학습 가능
- **결과** : 적은 학습 데이터만으로도 11개 데이터셋에서 handcrafted prompt 대비 우수한 few-shot classification 성능 달성



1. CoOp

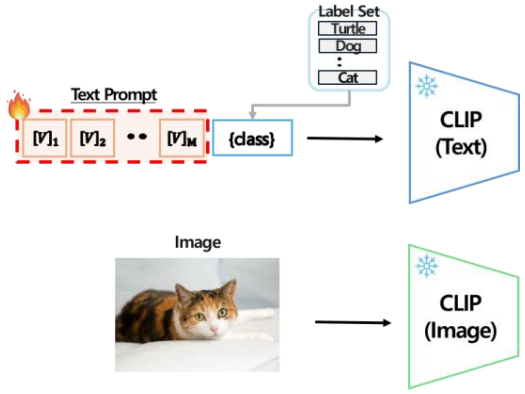
Context Optimization (CoOp)

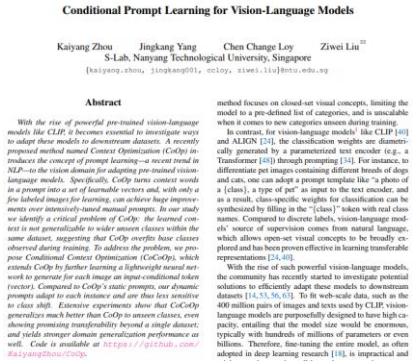
❖ Prompt Learning 의 발전과정

Text Prompt | **Multimodal Prompt** | **Multimodal Representation**

CoOp [IJCV 2022] → **CoCoOp** [CVPR 2022] → **MaPLe** [CVPR 2023] → **MMRL** [CVPR 2025]

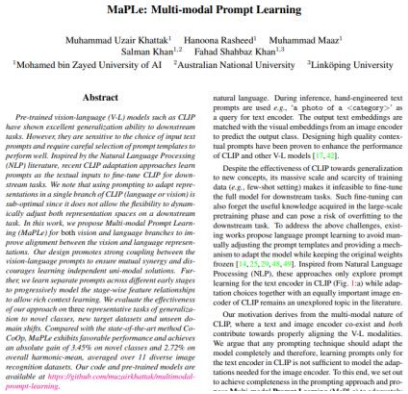
연속적인 Context vector 제안




Conditional Prompt Learning for Vision-Language Models

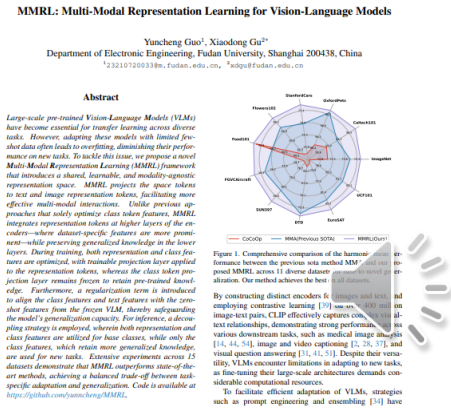
Kaiyang Zhou Jingkang Yang Chen Change Loy Ziwei Liu
S-Lab, Nanyang Technological University, Singapore

Abstract


MaPLe: Multi-modal Prompt Learning

Muhammad Uzair Khattak[1] Hanoona Rasheed[1] Muhammad Maaz[1]
Salman Khan[1,2] Fahad Shahbaz Khan[1,3]
[1]Mohamed bin Zayed University of AI [2]Australian National University [3]Linköping University

Abstract


MMRL: Multi-Modal Representation Learning for Vision-Language Models

Yuncheng Guo[1], Xiaodong Gu[2]*
Department of Electronic Engineering, Fudan University, Shanghai 200438, China

Abstract

2. CoCoOp

Conditional Context Optimization (CoCoOp)

❖ Prompt Learning 의 발전과정

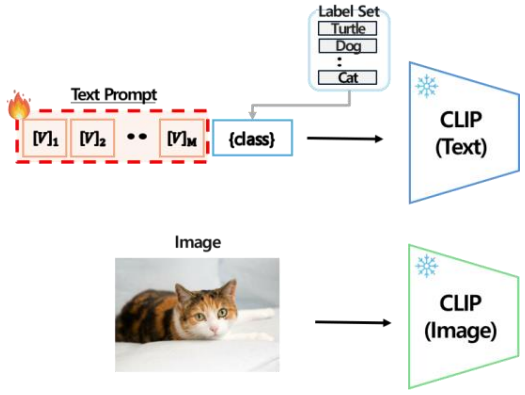
Text Prompt | **Multimodal Prompt** | **Multimodal Representation**

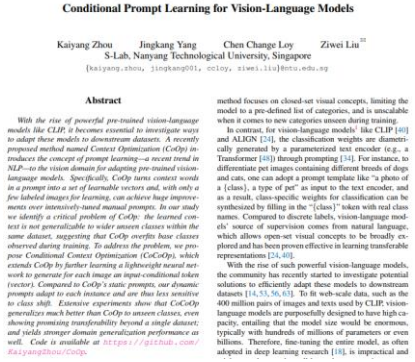
한계 :
Static Prompt 구조

CoOp [IJCV 2022] → **CoCoOp** [CVPR 2022] → **MaPLe** [CVPR 2023] → **MMRL** [CVPR 2025]

연속적인 Context vector 제안

조건화된 Prompt Learning

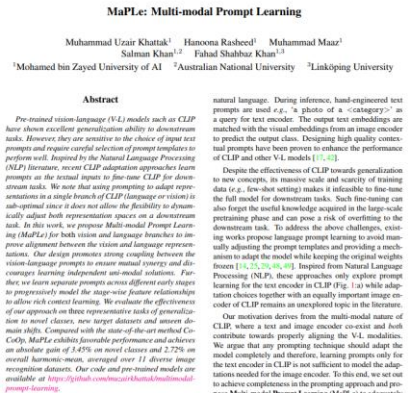



Conditional Prompt Learning for Vision-Language Models

Kaiyang Zhou Jingkang Yang Chen Change Loy Ziwei Liu
S-Lab, Nanyang Technological University, Singapore

Abstract

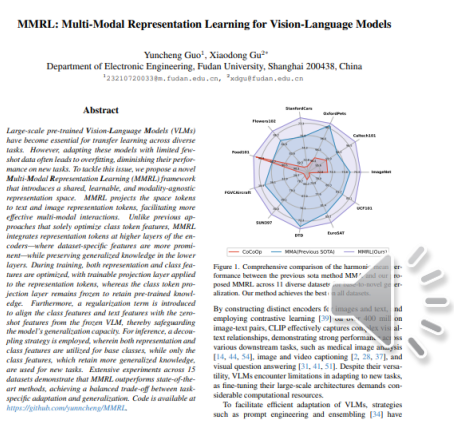
With the rise of powerful pre-trained vision-language models like CLIP, it becomes essential to investigate ways to adapt these models to downstream datasets. A recently proposed method named Context Optimization (CoOp) introduces the concept of prompt learning—a recent trend in NLP—to the vision domain for adapting pre-trained vision-language models. Specifically, CoOp turns context words in a prompt into a set of learnable vectors and, with only a few labeled images for learning, can achieve huge improvements over intensively-tuned manual prompts. In our study we identify a critical problem of CoOp: the learned context is not generalizable to wider unseen classes within the same dataset, suggesting that CoOp overfits base classes observed during training. To address the problem, we propose Conditional Context Optimization (CoCoOp), which extends CoOp by further learning a lightweight neural network to generate for each image an input-conditional token (vector). Compared to CoOp's static prompts, our dynamic prompts adapt to each instance and are thus less sensitive to class shift. Extensive experiments show that CoCoOp generalizes much better than CoOp to unseen classes, even showing promising transferability beyond a single dataset; and yields stronger domain generalization performance as well. Code is available at https://github.com/KaiyangZhou/CoOp.


MaPLe: Multi-modal Prompt Learning

Muhammad Uzair Khattak[1] Hanoona Rasheed[1] Muhammad Maaz[1]
Salman Khan[1,2] Fahad Shahbaz Khan[1,3]
[1]Mohamed bin Zayed University of AI [2]Australian National University [3]Linköping University

Abstract

Pre-trained vision-language (V-L) models such as CLIP have shown excellent generalization ability to downstream tasks. However, they are sensitive to the choice of input text prompts and require careful selection of prompt templates to perform well. Inspired by the Natural Language Processing (NLP) literature, recent CLIP adaptation approaches learn prompts as the textual inputs to fine-tune CLIP for downstream tasks. We note that using prompting to adapt representations in a single branch of CLIP (language or vision) is sub-optimal since it does not allow the flexibility to dynamically adjust both representation spaces on a downstream task. In this work, we propose Multi-modal Prompt Learning (MaPLe) for both vision and language branches to improve alignment between the vision and language representations. Our design promotes strong coupling between the vision-language prompts to ensure mutual synergy and discourages learning independent uni-modal solutions. Further, we learn separate prompts across different early stages to progressively model the stage-wise feature relationships to allow rich context learning. We evaluate the effectiveness of our approach on three representative tasks of generalization to novel classes, new target datasets and unseen domain shifts. Compared with the state-of-the-art method CoCoOp, MaPLe exhibits favorable performance and achieves an absolute gain of 3.45% on novel classes and 2.72% on overall harmonic-mean, averaged over 11 diverse image recognition datasets. Our code and pre-trained models are available at https://github.com/muzairkhattak/multimodal-prompt-learning.


MMRL: Multi-Modal Representation Learning for Vision-Language Models

Yuncheng Guo[1], Xiaodong Gu[2*]
Department of Electronic Engineering, Fudan University, Shanghai 200438, China

Abstract

Large-scale pre-trained Vision-Language Models (VLMs) have become essential for transfer learning across diverse tasks. However, adapting these models with limited few-shot data often leads to overfitting, diminishing their performance on new tasks. To tackle this issue, we propose a novel Multi-Modal Representation Learning (MMRL) framework that introduces a shared, learnable, and modality-agnostic representation space. MMRL projects the space tokens to text and image representation tokens, facilitating more effective multi-modal interactions. Unlike previous approaches that solely optimize class token features, MMRL integrates representation tokens at higher layers of the encoders—where dataset-specific features are more prominent—while preserving generalized knowledge in the lower layers. During training, both representation and class features are optimized, with trainable projection layer applied to the representation tokens, whereas the class token projection layer remains frozen to retain pre-trained knowledge. Furthermore, a regularization term is introduced to align the class features and text features with the zero-shot features from the frozen VLM, thereby safeguarding the model's generalization capacity. For inference, a decoupling strategy is employed, wherein both representation and class features are utilized for base classes, while only the class features, which retain more generalized knowledge, are used for new tasks. Extensive experiments across 15 datasets demonstrate that MMRL outperforms state-of-the-art methods, achieving a balanced trade-off between task-specific adaptation and generalization. Code is available at https://github.com/yunncheng/MMRL.

2. CoCoOp

Conditional Context Optimization (CoCoOp)

❖ CoOp 은 어떤 한계가 있었을까?

- **한계점** : Base class에는 강하지만, novel class에서는 성능이 크게 하락하는 Base-to-New Generalization 문제 발생
- Base class에 최적화된 prompt가 특정 클래스 분포에 과적합되어, unseen class에 대한 일반화 성능이 제한됨

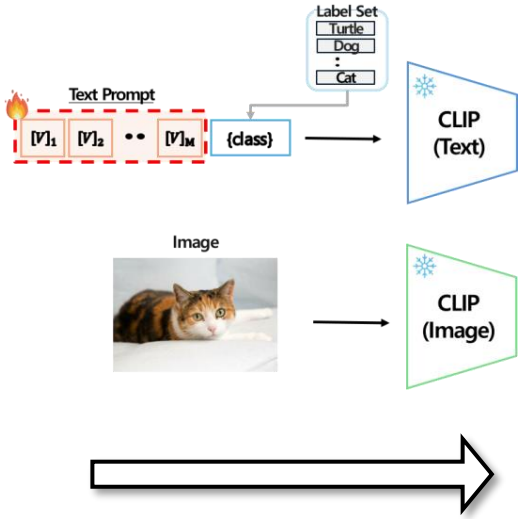
CoOp
[IJCV 2022]

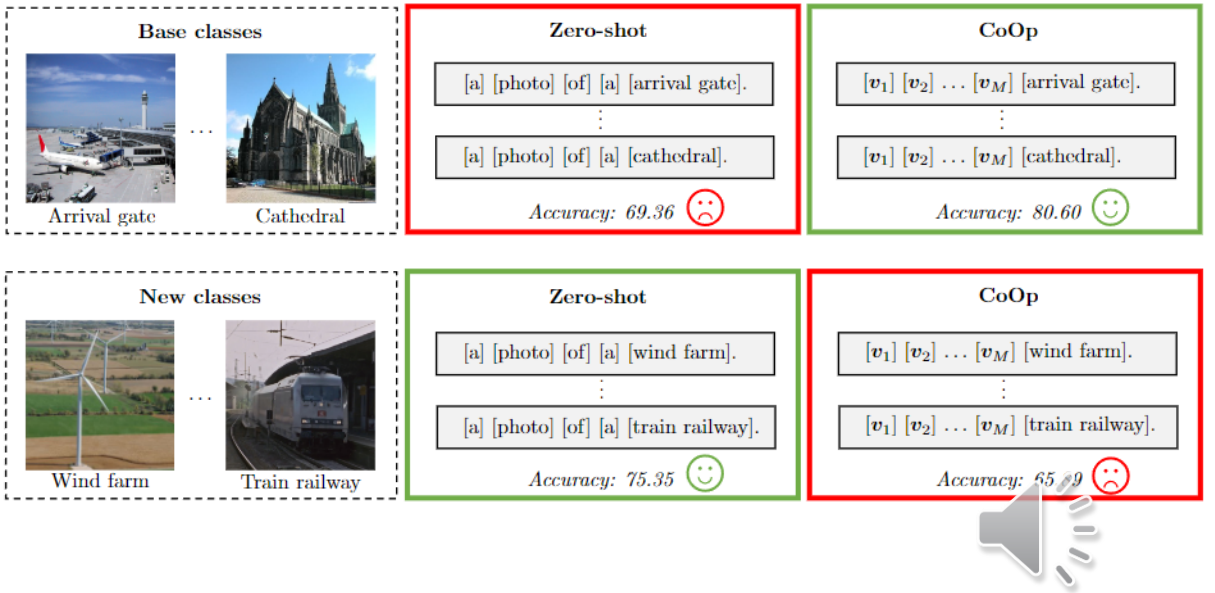
Learning to Prompt for Vision-Language Models

Kaiyang Zhou · Jingkang Yang · Chen Change Loy · Ziwei Liu

Received: date / Accepted: date

Abstract Large pre-trained vision-language models like CLIP have shown great potential in learning representations that are transferable across a wide range of downstream tasks. Different from the traditional representation learning that is based mostly on discretized labels, vision-language pre-training aligns images and texts in a common feature space, which allows zero-shot transfer to a downstream task via *prompting*, i.e., classification weights are synthesized from natural language describing classes of interest. In this work, we show that a major challenge for deploying such models in practice is prompt engineering, which requires domain expertise and is extremely time-consuming—implementations of CoOp: unified context and class-specific context. Through extensive experiments on 11 datasets, we demonstrate that CoOp requires as few as one or two shots to beat hand-crafted prompts with a decent margin and is able to gain significant improvements over prompt engineering with more shots, e.g., with 16 shots the average gain is around 15% (with the highest reaching over 45%). Despite being a learning-based approach, CoOp achieves superb domain generalization performance compared with the zero-shot model using hand-crafted prompts.





2. CoCoOp

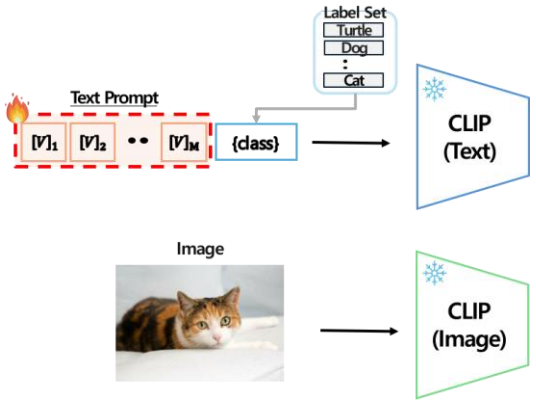
Conditional Context Optimization (CoCoOp)

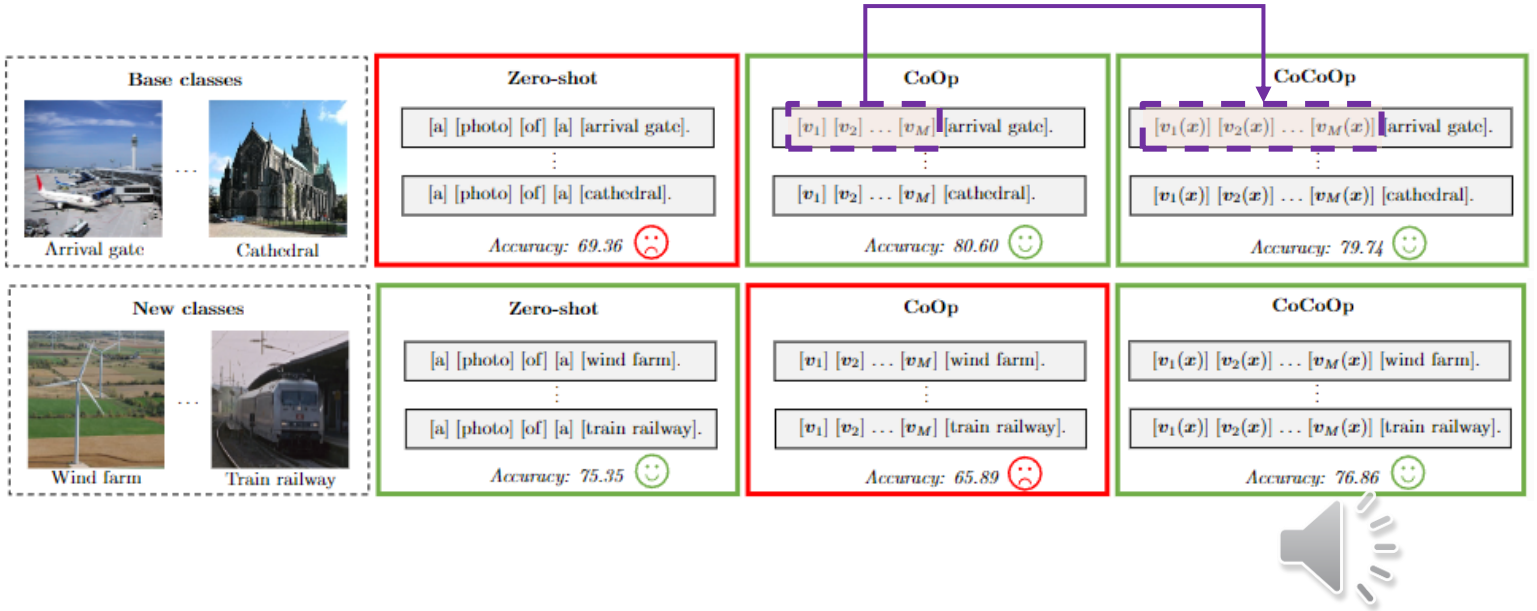
❖ CoOp 은 어떤 한계가 있었을까?

- **한계점** : Base class에는 강하지만, novel class에서는 성능이 크게 하락하는 Base-to-New generalization 문제 발생
- Base class에 최적화된 prompt가 특정 클래스 분포에 과적합되어, unseen class에 대한 일반화 성능이 제한됨

  → **CoCoOp** [1] : CoOp 의 static prompt 구조와 달리, **입력되는 이미지에 조건화된 프롬프트를 학습하는 것을 제안**

CoOp
[IJCV 2022] → **CoCoOp** [CVPR 2022]





[1] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual Prompt Tuning. In Proceedings of the European Conference on Computer Vision (ECCV), 2022.

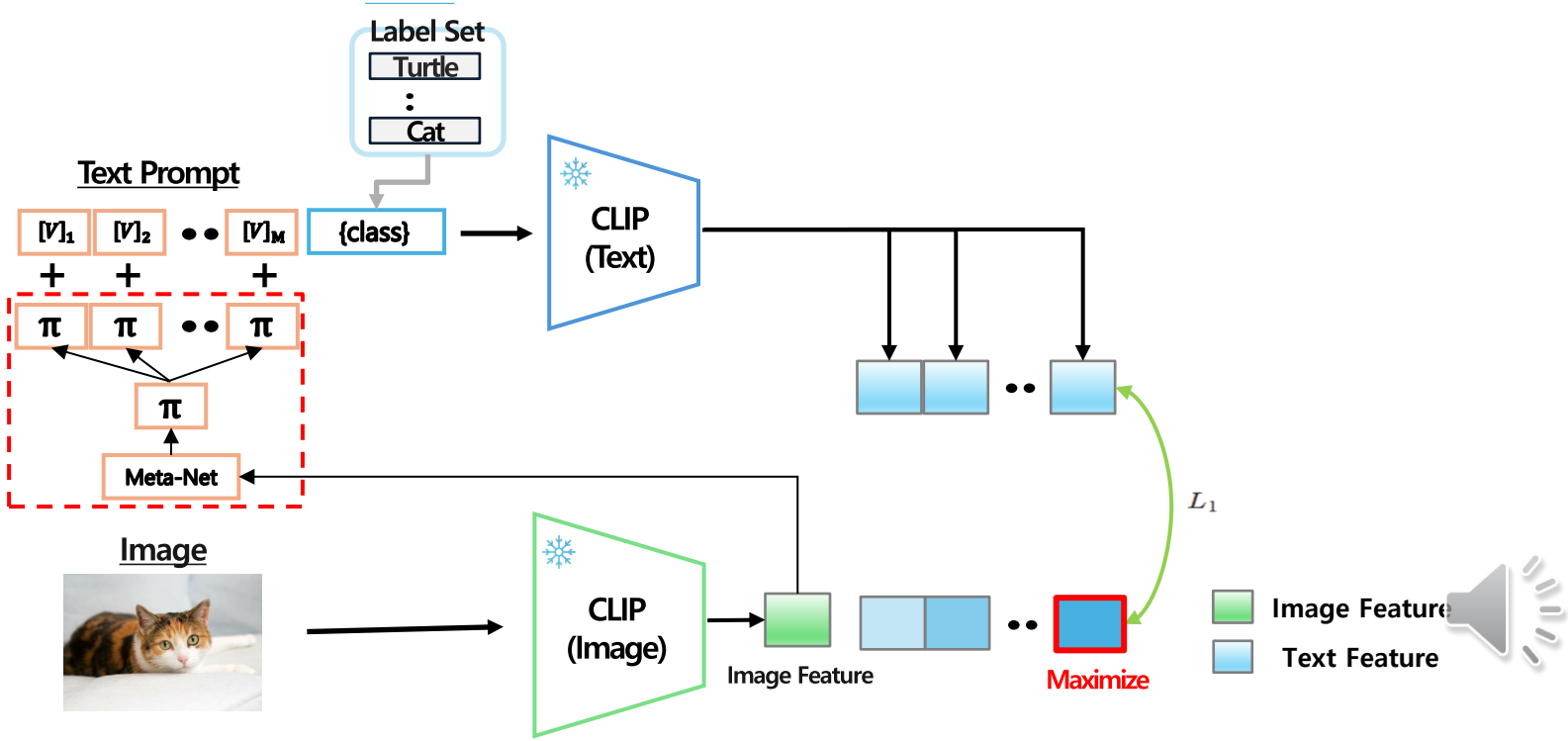
2. CoCoOp

Conditional Context Optimization (CoCoOp)

❖ Conditional Prompt Learning for Vision-Language Models [CVPR 2022]

- **주요 기여** : 입력 이미지의 Visual feature에 조건화된 prompt를 생성
- **Meta-Net:** Linear–ReLU–Linear로 구성된 2-layer bottleneck network

  ⇒ Image feature를 하나의 conditional token π 로 변환한 뒤, 이를 모든 context vector에 더해 $v_m(x) = v_m + \pi$ 로 구성



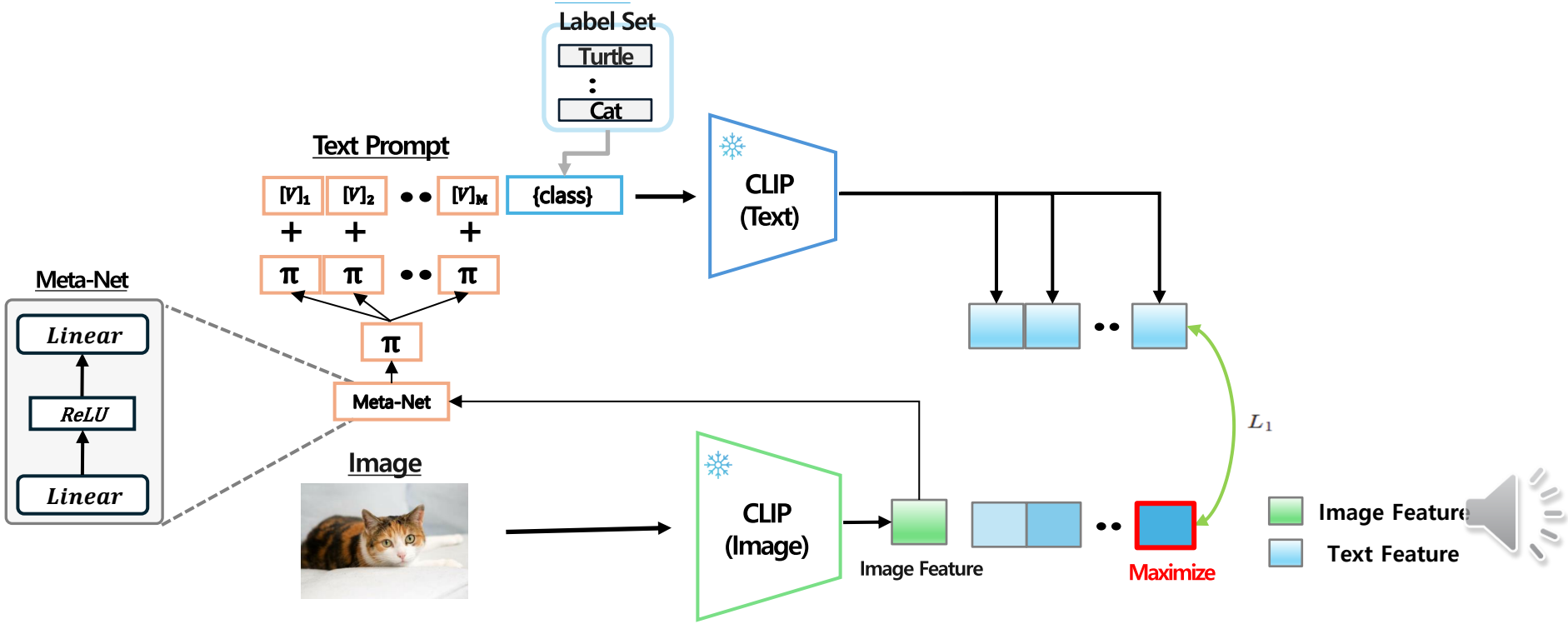
2. CoCoOp

Conditional Context Optimization (CoCoOp)

❖ Conditional Prompt Learning for Vision-Language Models [CVPR 2022]

- **주요 기여** : 입력 이미지의 Visual feature에 조건화된 prompt를 생성
- **Meta-Net:** Linear–ReLU–Linear로 구성된 2-layer bottleneck network

  ⇒ Image feature를 하나의 conditional token π 로 변환한 뒤, 이를 모든 context vector에 더해 $v_m(x) = v_m + \pi$ 로 구성



2. CoCoOp

Conditional Context Optimization (CoCoOp)

❖ Evaluation Setting

✓ 평가 데이터셋

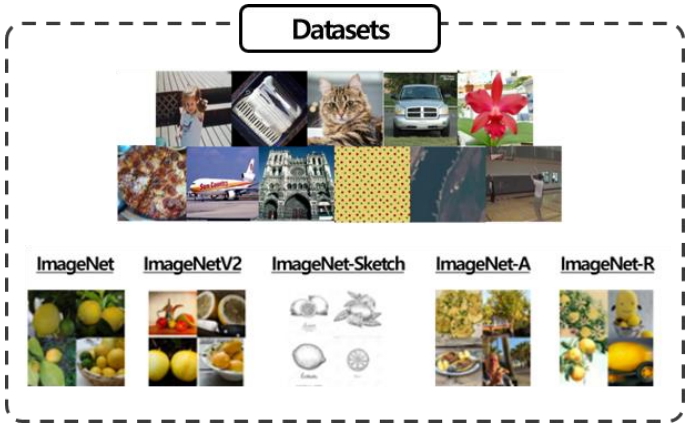
- ImageNet, Caltech101, OxfordPets, StanfordCars, Flowers102, Food101, FGVCAircraft, SUN397, DTD, EuroSAT, UCF101
- ImageNet, ImageNetV2, ImageNet-Sketch, ImageNet-A, ImageNet-R

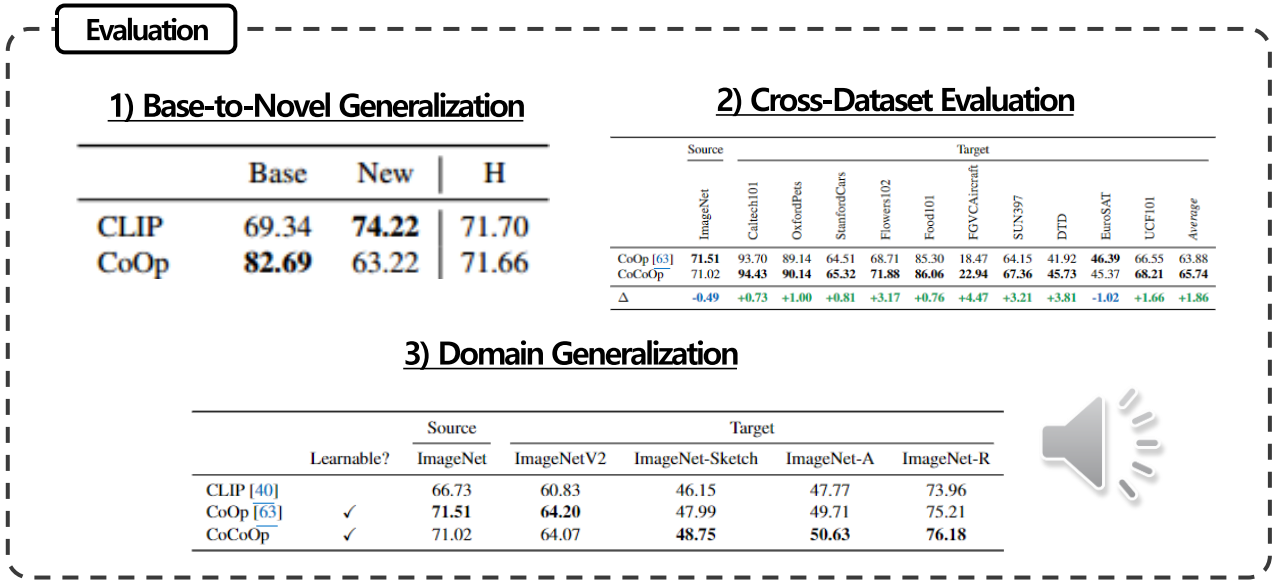
✓ **Generalization Performance**

(1) Base-to-New Generalization (B2N)

(2) Cross-Dataset Evaluation (CD)

(3) Domain Generalization (DG)





	Base	New	H
CLIP	69.34	**74.22**	71.70
CoOp	**82.69**	63.22	71.66

	Source	Target									
	ImageNet	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101	*Average*
CoOp [63]	**71.51**	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	**46.39**	66.55	63.88
CoCoOp	71.02	**94.43**	**90.14**	**65.32**	**71.88**	**86.06**	**22.94**	**67.36**	**45.73**	45.37	**68.21**	**65.74**
Δ	-0.49	+0.73	+1.00	+0.81	+3.17	+0.76	+4.47	+3.21	+3.81	-1.02	+1.66	+1.86

	Learnable?	Source	Target			
		ImageNet	ImageNetV2	ImageNet-Sketch	ImageNet-A	ImageNet-R
CLIP [40]		66.73	60.83	46.15	47.77	73.96
CoOp [63]	✓	**71.51**	**64.20**	47.99	49.71	75.21
CoCoOp	✓	71.02	64.07	**48.75**	**50.63**	**76.18**

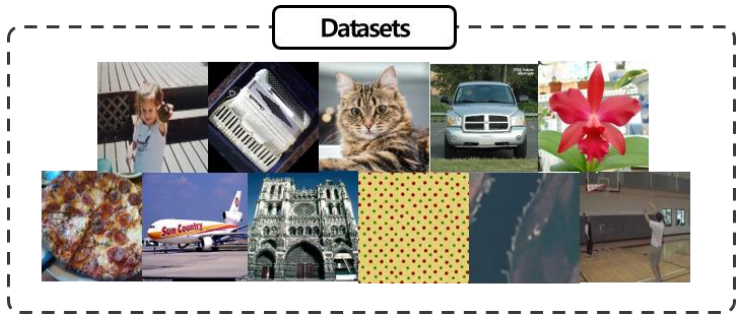
2. CoCoOp

Conditional Context Optimization (CoCoOp)

❖ Evaluation Setting

(1) Base-to-New Generalization (B2N)

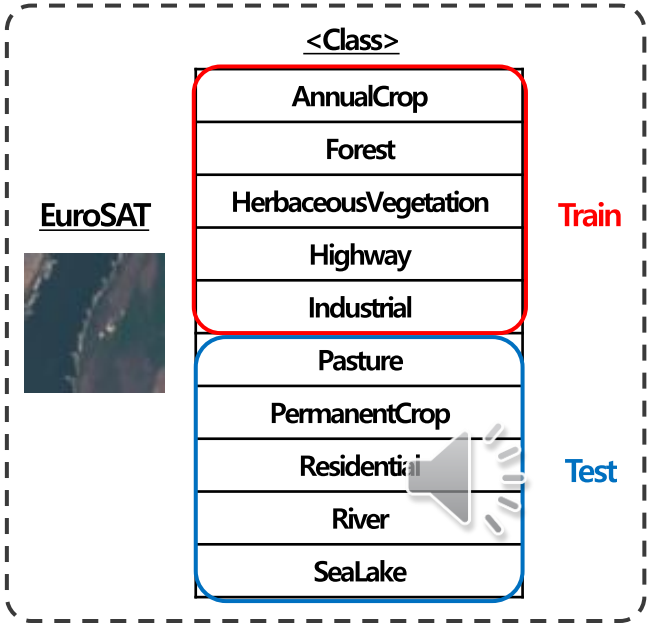
- **목적 : 학습 클래스 적응과 새로운 클래스 일반화의 균형**을 평가
- 데이터셋 내 클래스를 Base/New로 분할하고, Base class의 16-shot 데이터로 prompt 학습
- **Base Accuracy**는 학습 클래스 적응 성능, **New Accuracy**는 unseen class 일반화 성능을 의미하며, **Harmonic Mean**으로 두 성능의 균형을 평가



1) Base-to-New Generalization

	Train: Base	Test: New	H
CLIP	69.34	**74.22**	71.70
CoOp	**82.69**	63.22	71.66

$$H = \frac{2ab}{a + b}$$



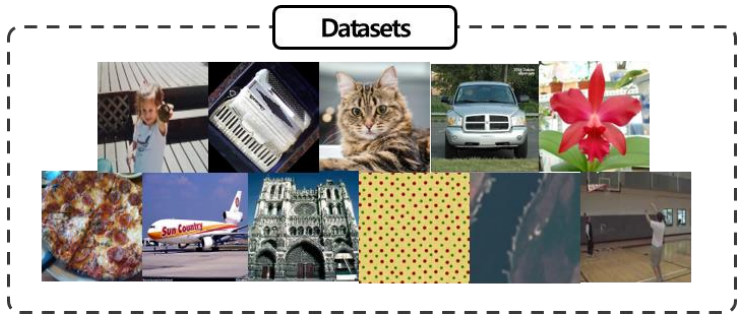
2. CoCoOp

Conditional Context Optimization (CoCoOp)

❖ Evaluation Setting

(2) Cross-Dataset Evaluation (CD)

- **목적 :** 학습 데이터와 클래스 및 도메인 특성이 다른 Target Dataset에서도 성능이 유지되는지 평가하여 **데이터셋 간 전이 가능성**을 검증
- ImageNet에서 클래스당 16장으로 학습한 prompt를 **추가 학습 없이 나머지 10개 데이터셋에 직접 적용**



2) Cross-Dataset Evaluation

	Train	Test										
	Source	Target										
	ImageNet	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101	*Average*
CoOp [63]	**71.51**	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	**46.39**	66.55	63.88
CoCoOp	71.02	**94.43**	**90.14**	**65.32**	**71.88**	**86.06**	**22.94**	**67.36**	**45.73**	45.37	**68.21**	**65.74**
Δ	-0.49	+0.73	+1.00	+0.81	+3.17	+0.76	+4.47	+3.21	+3.81	-1.02	+1.66	+1.86

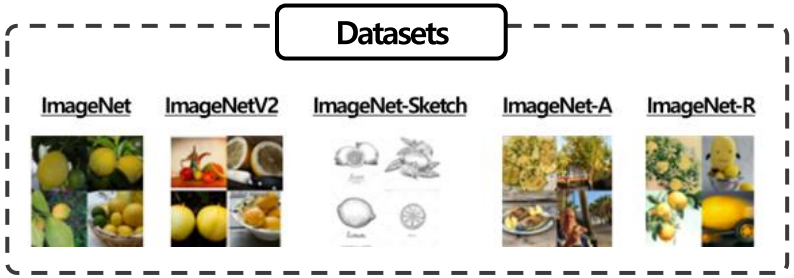
2. CoCoOp

Conditional Context Optimization (CoCoOp)

❖ Evaluation Setting

(3) Domain Generalization (DG)

- **목적 :** 클래스 의미는 유지되지만 수집 시점, 스타일, 난이도, 표현 방식이 달라진 환경에서 **도메인 강건성을 평가**
- ImageNet에서 학습한 prompt를 ImageNetV2, ImageNet-Sketch, ImageNet-A, ImageNet-R에 **추가 학습 없이 적용**



3) Domain Generalization

		Train	Test			
		Source	Target			
	Learnable?	ImageNet	ImageNetV2	ImageNet-Sketch	ImageNet-A	ImageNet-R
CLIP [40]		66.73	60.83	46.15	47.77	73.96
CoOp [63]	✓	**71.51**	**64.20**	47.99	49.71	75.21
CoCoOp	✓	71.02	64.07	**48.75**	**50.63**	**76.18**

2. CoCoOp

Conditional Context Optimization (CoCoOp)

❖ Evaluation Setting

✓ **Model Efficiency 평가**

- Prompt Learning은 적은 파라미터로 VLM을 효율적으로 적응시키는 것이 핵심 목적

  → 성능 향상이 모델 규모가 아닌 방법론 자체에서 기인하는지 확인하기 위해 Parameter 및 Computational Efficiency를 평가

- 일반화 성능과 학습·추론 비용 사이의 Trade-Off를 비교하여 실제 적용 가능성을 검증

2. CoCoOp

Conditional Context Optimization (CoCoOp)

❖ Result

(1) Base-to-New Generalization (B2N)

- CoCoOp은 CoOp 대비 Base accuracy는 **82.69 → 80.47 로** 소폭 감소했지만, New accuracy는 **63.22 → 71.69** 로 크게 향상
- 11개 데이터셋 중 5개에서 New accuracy가 10%p 이상 향상되었으며, Base 성능 감소보다 New 성능 향상이 전반적으로 더 크게 나타남
- CoCoOp의 일반화 성능 강화 특성으로 인해 Base acc는 대부분 3%p 미만으로 소폭 하락했지만, New class에서의 큰 성능 향상을 보임

1) Base-to-New Generalization

(a) **Average over 11 datasets.**

	Base	New	H
CLIP	69.34	**74.22**	71.70
CoOp	**82.69**	63.22	71.66
CoCoOp	80.47	71.69	**75.83**

(b) ImageNet.

	Base	New	H
CLIP	72.43	68.14	70.22
CoOp	**76.47**	67.88	71.92
CoCoOp	75.98	**70.43**	**73.10**

(c) Caltech101.

	Base	New	H
CLIP	96.84	**94.00**	95.40
CoOp	**98.00**	89.81	93.73
CoCoOp	97.96	93.81	**95.84**

(d) OxfordPets.

	Base	New	H
CLIP	91.17	97.26	94.12
CoOp	93.67	95.29	94.47
CoCoOp	**95.20**	**97.69**	**96.43**

(e) StanfordCars.

	Base	New	H
CLIP	63.37	**74.89**	68.65
CoOp	**78.12**	60.40	68.13
CoCoOp	70.49	73.59	**72.01**

(f) Flowers102.

	Base	New	H
CLIP	72.08	**77.80**	74.83
CoOp	**97.60**	59.67	74.06
CoCoOp	94.87	71.75	**81.71**

(g) Food101.

	Base	New	H
CLIP	90.10	91.22	90.66
CoOp	88.33	82.26	85.19
CoCoOp	**90.70**	**91.29**	**90.99**

(h) FGVCAircraft.

	Base	New	H
CLIP	27.19	**36.29**	**31.09**
CoOp	**40.44**	22.30	28.75
CoCoOp	33.41	23.71	27.74

(i) SUN397.

	Base	New	H
CLIP	69.36	75.35	72.23
CoOp	**80.60**	65.89	72.51
CoCoOp	79.74	**76.86**	**78.27**

(j) DTD.

	Base	New	H
CLIP	53.24	**59.90**	56.37
CoOp	**79.44**	41.18	54.24
CoCoOp	77.01	56.00	**64.85**

(k) EuroSAT.

	Base	New	H
CLIP	56.48	**64.05**	60.03
CoOp	**92.19**	54.74	68.69
CoCoOp	87.49	60.04	**71.21**

(l) UCF101.

	Base	New	H
CLIP	70.53	**77.50**	73.85
CoOp	**84.69**	56.05	67.46
CoCoOp	82.33	73.45	**77.64**

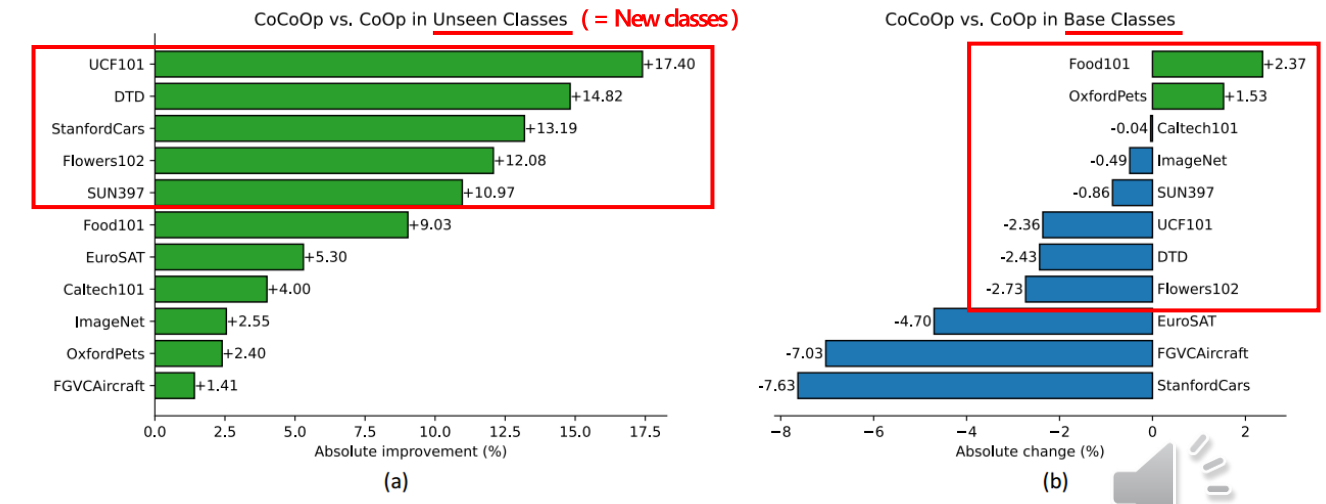


Figure 3. **Comprehensive comparisons of CoCoOp and CoOp in the base-to-new generalization setting**. (a) CoCoOp is able to gain consistent improvements over CoOp in unseen classes on all datasets. (b) CoCoOp's declines in base accuracy are mostly under 3%, which are far outweighed by the gains in generalization.

2. CoCoOp

Conditional Context Optimization (CoCoOp)

❖ Result

(2) Cross-Dataset Evaluation (CD)

- ImageNet에서 학습한 prompt를 10개 target dataset에 적용했을 때, 평균 정확도가 **63.88 → 65.74**로 향상
- 대부분의 target dataset에서 CoOp을 능가했으며, 특히 **FGVCAircraft, DTD 등 전문화된 데이터셋에서 전이 성능 개선**

(3) Domain Generalization (DG)

- ImageNetV2에서는 CoOp보다 성능이 소폭 낮았지만, **ImageNet-Sketch, ImageNet-A, ImageNet-R에서는 더 높은 성능**
- 입력 이미지에 조건화된 Prompt가 정적인 Prompt보다 **분포 변화와 표현 방식 변화에 더 강건함**을 확인

2) Cross-Dataset Evaluation

	Source	Target										
	ImageNet	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101	*Average*
CoOp [63]	**71.51**	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	**46.39**	66.55	63.88
CoCoOp	71.02	**94.43**	**90.14**	**65.32**	**71.88**	**86.06**	**22.94**	**67.36**	**45.73**	45.37	**68.21**	**65.74**
Δ	-0.49	+0.73	+1.00	+0.81	+3.17	+0.76	+4.47	+3.21	+3.81	-1.02	+1.66	+1.86

3) Domain Generalization

	Learnable?	Source	Target			
		ImageNet	ImageNetV2	ImageNet-Sketch	ImageNet-A	ImageNet-R
CLIP [40]		66.73	60.83	46.15	47.77	73.96
CoOp [63]	✓	**71.51**	**64.20**	47.99	49.71	75.21
CoCoOp	✓	71.02	64.07	**48.75**	**50.63**	**76.18**

2. CoCoOp

Conditional Context Optimization (CoCoOp)

❖ Model Efficiency

(1) Params

- Meta-Net을 제거한 뒤 CoOp의 Context Token 수를 늘려, CoOp과 CoCoOp의 파라미터 규모를 유사하게 설정
- 유사한 파라미터 수의 Large CoOp보다 CoCoOp의 New/H 성능이 높음

→ 모델 성능 향상은 파라미터 증가가 아닌 **conditional prompting 효과**란 것을 보임

Table 5. CoCoOp (last row) vs a bigger CoOp on ImageNet.

Model	# params	Base	New	H
CoOp (ctx=4)	2,048	**76.47**	67.88	71.92
CoOp (ctx=60)	30,720	76.16	65.34	70.34
CoOp (ctx=4) + Meta-Net	34,816	75.98	**70.43**	**73.10**

(2) Computational Cost

- **한계 :** CoCoOp은 입력 이미지마다 class별 Text Feature를 새로 계산해야 하므로, 학습 속도가 느려지고 GPU memory 사용량이 증가함

2. CoCoOp

Conditional Context Optimization (CoCoOp)

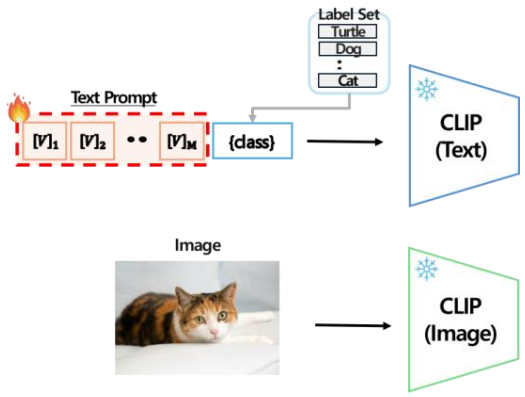
❖ Prompt Learning 의 발전과정

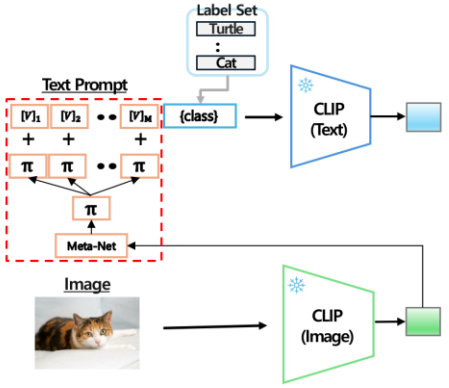
Text Prompt | **Multimodal Prompt** | **Multimodal Representation**

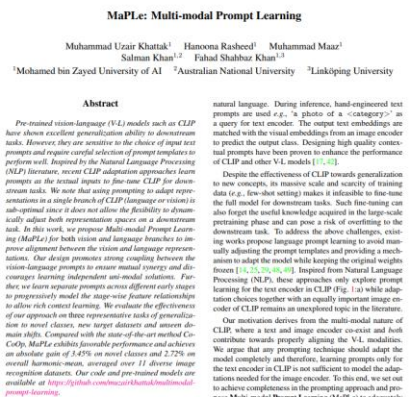
CoOp [IJCV 2022] → **CoCoOp** [CVPR 2022] → **MaPLe** [CVPR 2023] → **MMRL** [CVPR 2025]

연속적인 Context vector 제안

조건화된 Prompt Learning 제안







MaPLe: Multi-modal Prompt Learning

Muhammad Uzair Khattak[1] Hanoona Rasheed[1] Muhammad Maaz[1]
Salman Khan[1,2] Fahad Shahbaz Khan[1,3]
[1]Mohamed bin Zayed University of AI [2]Australian National University [3]Linköping University

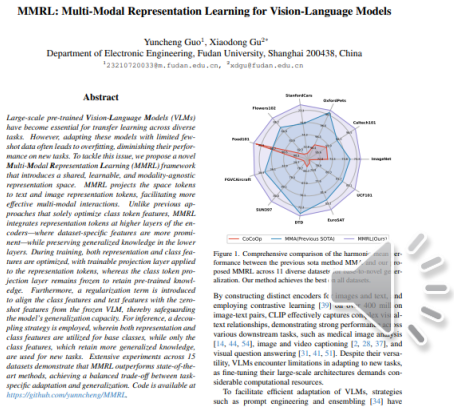
Abstract

Pre-trained vision-language (V-L) models such as CLIP have shown excellent generalization ability to downstream tasks. However, they are sensitive to the choice of input text prompts and require careful selection of prompt templates to perform well. Inspired by the Natural Language Processing (NLP) literature, recent CLIP adaptation approaches learn prompts as the textual inputs to fine-tune CLIP for downstream tasks. We note that using prompting to adapt representations in a single branch of CLIP (language or vision) is sub-optimal since it does not allow the flexibility to dynamically adjust both representation spaces on a downstream task. In this work, we propose Multi-modal Prompt Learning (MaPLe) for both vision and language branches to improve alignment between the vision and language representations. Our design promotes strong coupling between the vision-language prompts to ensure mutual synergy and discourages learning independent uni-modal solutions. Further, we learn separate prompts across different early stages to progressively model the stage-wise feature relationships to allow rich context learning. We evaluate the effectiveness of our approach on three representative tasks of generalization to novel classes, new target datasets and unseen domain shifts. Compared with the state-of-the-art method CoCoOp, MaPLe exhibits favorable performance and achieves an absolute gain of 3.45% on novel classes and 2.72% on overall harmonic-mean, averaged over 11 diverse image recognition datasets. Our code and pre-trained models are available at https://github.com/muzairkhattak/multimodal-prompt-learning.

natural language. During inference, hand-engineered text prompts are used *e.g.*, 'a photo of a <<category>>' as a query for text encoder. The output text embeddings are matched with the visual embeddings from an image encoder to predict the output class. Designing high quality contextual prompts have been proven to enhance the performance of CLIP and other V-L models [17, 42].

Despite the effectiveness of CLIP towards generalization to new concepts, its massive scale and scarcity of training data (e.g., few-shot setting) makes it infeasible to fine-tune the full model for downstream tasks. Such fine-tuning can also forget the useful knowledge acquired in the large-scale pretraining phase and can pose a risk of overfitting to the downstream task. To address the above challenges, existing works propose language prompt learning to avoid manually adjusting the prompt templates and providing a mechanism to adapt the model while keeping the original weights frozen [14, 25, 29, 48, 49]. Inspired from Natural Language Processing (NLP), these approaches only explore prompt learning for the text encoder in CLIP (Fig. 1a) while adaptation choices together with an equally important image encoder of CLIP remains an unexplored topic in the literature.

Our motivation derives from the multi-modal nature of CLIP, where a text and image encoder co-exist and both contribute towards properly aligning the V-L modalities. We argue that any prompting technique should adapt the model completely and therefore, learning prompts only for the text encoder in CLIP is not sufficient to model the adaptations needed for the image encoder. To this end, we set out to achieve completeness in the prompting approach and pro-



MMRL: Multi-Modal Representation Learning for Vision-Language Models

Yuncheng Guo[1], Xiaodong Gu[2*]
Department of Electronic Engineering, Fudan University, Shanghai 200438, China
[1]23210720033@m.fudan.edu.cn, [2*]xdgu@fudan.edu.cn

Abstract

Large-scale pre-trained Vision-Language Models (VLMs) have become essential for transfer learning across diverse tasks. However, adapting these models with limited few-shot data often leads to overfitting, diminishing their performance on new tasks. To tackle this issue, we propose a novel Multi-Modal Representation Learning (MMRL) framework that introduces a shared, learnable, and modality-agnostic representation space. MMRL projects the space tokens to text and image representation tokens, facilitating more effective multi-modal interactions. Unlike previous approaches that solely optimize class token features, MMRL integrates representation tokens at higher layers of the encoders—where dataset-specific features are more prominent—while preserving generalized knowledge in the lower layers. During training, both representation and class features are optimized, with trainable projection layer applied to the representation tokens, whereas the class token projection layer remains frozen to retain pre-trained knowledge. Furthermore, a regularization term is introduced to align the class features and text features with the zero-shot features from the frozen VLM, thereby safeguarding the model's generalization capacity. For inference, a decoupling strategy is employed, wherein both representation and class features are utilized for base classes, while only the class features, which retain more generalized knowledge, are used for new tasks. Extensive experiments across 15 datasets demonstrate that MMRL outperforms state-of-the-art methods, achieving a balanced trade-off between task-specific adaptation and generalization. Code is available at https://github.com/yunncheng/MMRL.

Figure 1. Comprehensive comparison of the harmonic mean performance between the previous sota method MMA and our proposed MMRL across 11 diverse datasets. MMRL++ is a novel parameter-efficient variation. Our method achieves the best trade-off across all datasets.

By constructing distinct encoders for images and text, and employing contrastive learning [39], we aim to align the two types of features, ensuring that similar image-text pairs, CLIP effectively captures complex, non-linear text relationships, demonstrating strong performance across various downstream tasks, such as medical image analysis [14, 44, 54], image and video captioning [2, 28, 37], and visual question answering [31, 41, 51]. Despite their versatility, VLMs encounter limitations in adapting to new tasks, as fine-tuning their large-scale architectures demands considerable computational resources.

To facilitate efficient adaptation of VLMs, strategies such as prompt engineering and ensembling [34] have

3. MaPLe

Multimodal Prompt Learning (MaPLe)

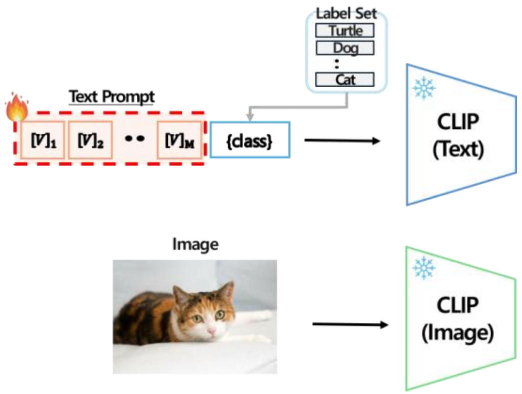
❖ Prompt Learning 의 발전과정

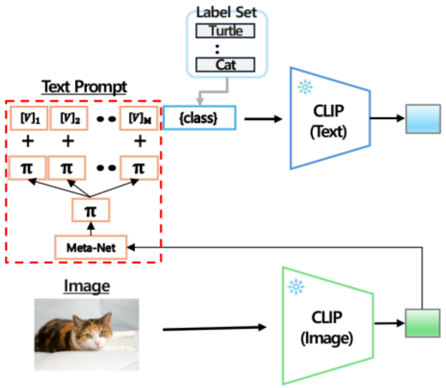
Text Prompt | **Multimodal Prompt** | **Multimodal Representation**

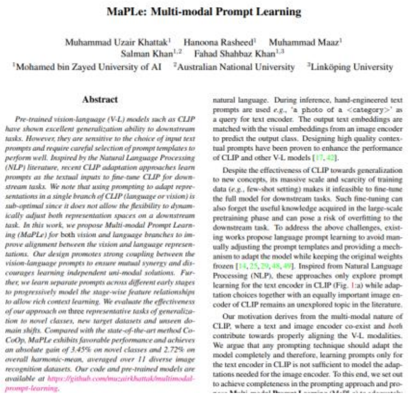
한계 :
Uni-Modal Prompting

CoOp [IJCV 2022] → **CoCoOp** [CVPR 2022] → **MaPLe** [CVPR 2023] → **MMRL** [CVPR 2025]

- CoOp: 연속적인 Context vector 제안
- CoCoOp: 조건화된 Prompt Learning 제안
- MaPLe: Multimodal Prompt Learning



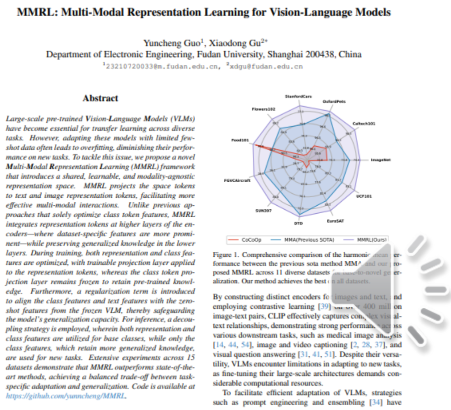



MaPLe: Multi-modal Prompt Learning

Muhammad Uzair Khattak[1] Hanoona Rasheed[1] Muhammad Maaz[1] Salman Khan[1,2] Fahad Shahbaz Khan[1,3]

[1]Mohamed bin Zayed University of AI [2]Australian National University [3]Linköping University

Abstract


MMRL: Multi-Modal Representation Learning for Vision-Language Models

Yuncheng Guo[1], Xiaodong Gu[2*]

Department of Electronic Engineering, Fudan University, Shanghai 200438, China

Abstract

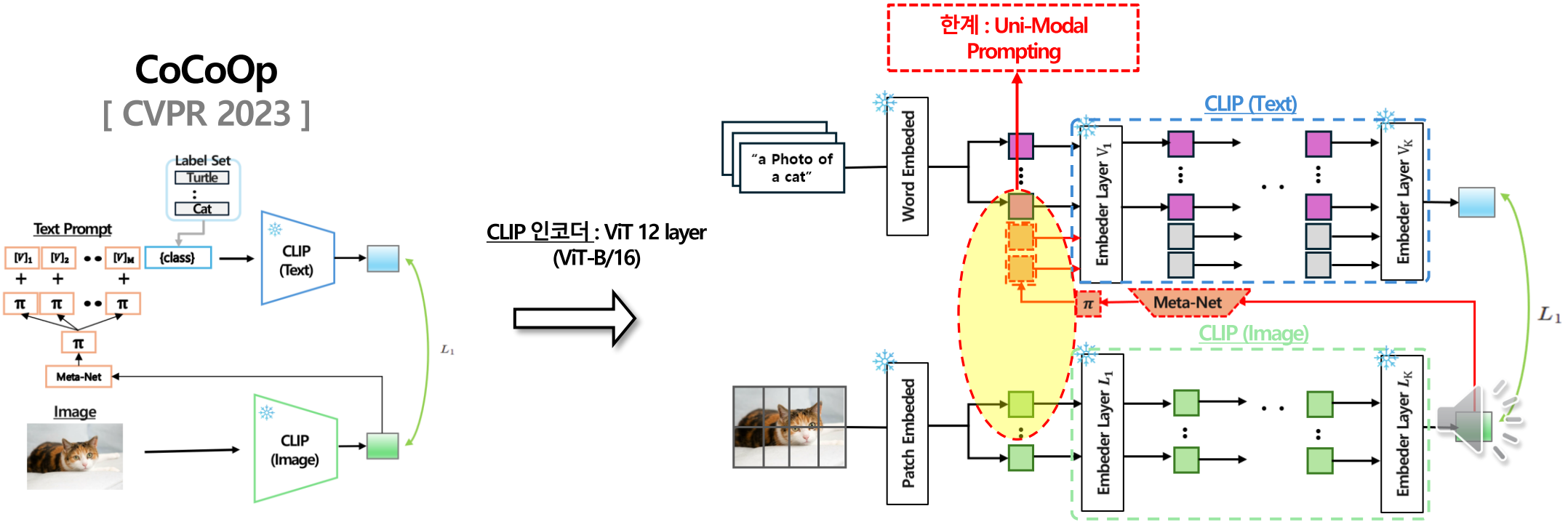
3. MaPLe

Multimodal Prompt Learning (MaPLe)

❖ CoCoOp 은 어떤 한계가 있었을까?

- **한계** : Text branch만 조정되며 두 modality가 서로 다른 방향으로 최적화됨으로써 이로 인해 CLIP의 Vision-Language 정렬이 약화될 가능성이 존재함

  → **MaPLe** [4] : Text와 Visual branch를 함께 조정하는 **Multimodal Prompt Learning 제안**



3. MaPLe

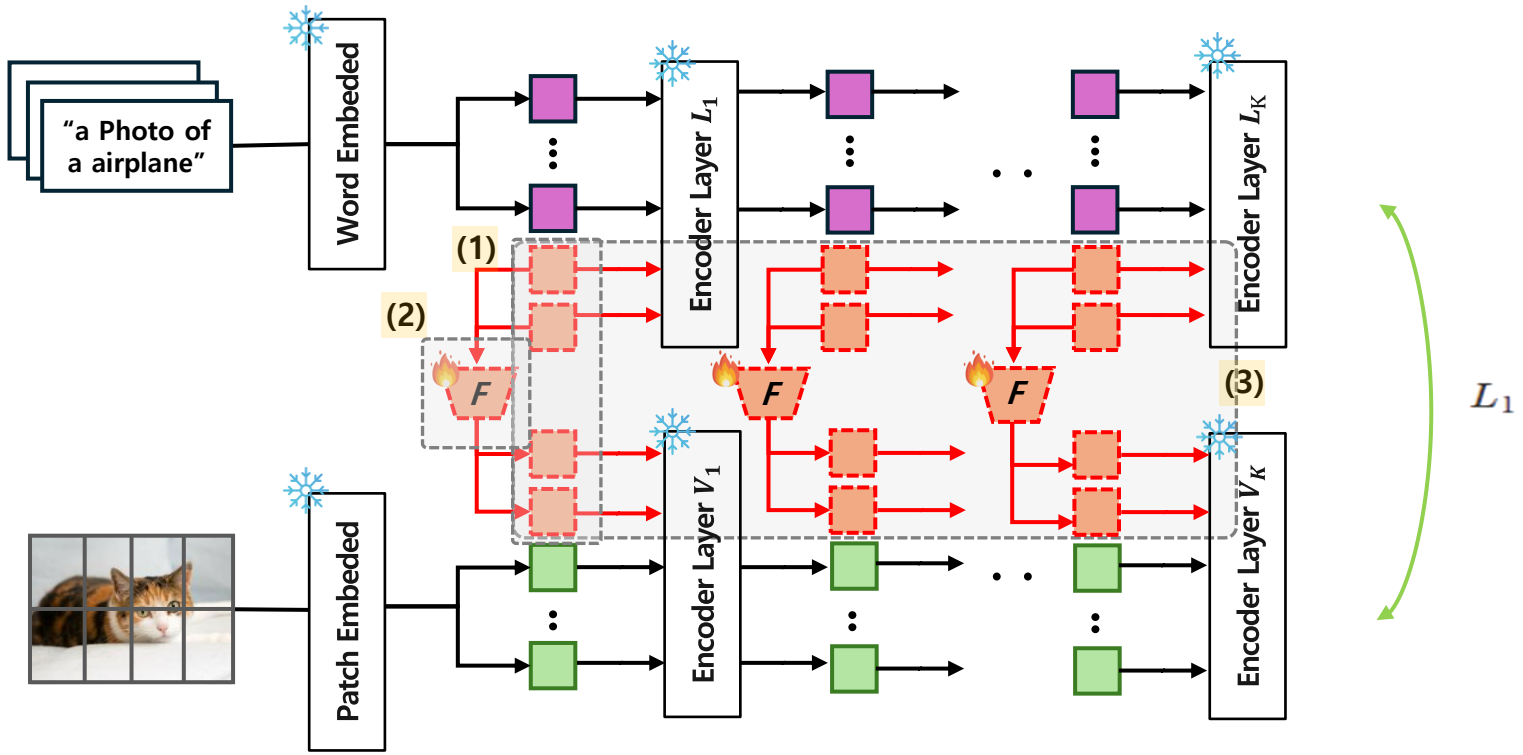
Multimodal Prompt Learning (MaPLe)

❖ MaPLe: Multimodal Prompt Learning [CVPR 2023]

(1) 제안 1 : Multimodal Prompting

(2) 제안 2 : V-L coupling function

(3) 제안 3 : Deep Prompting



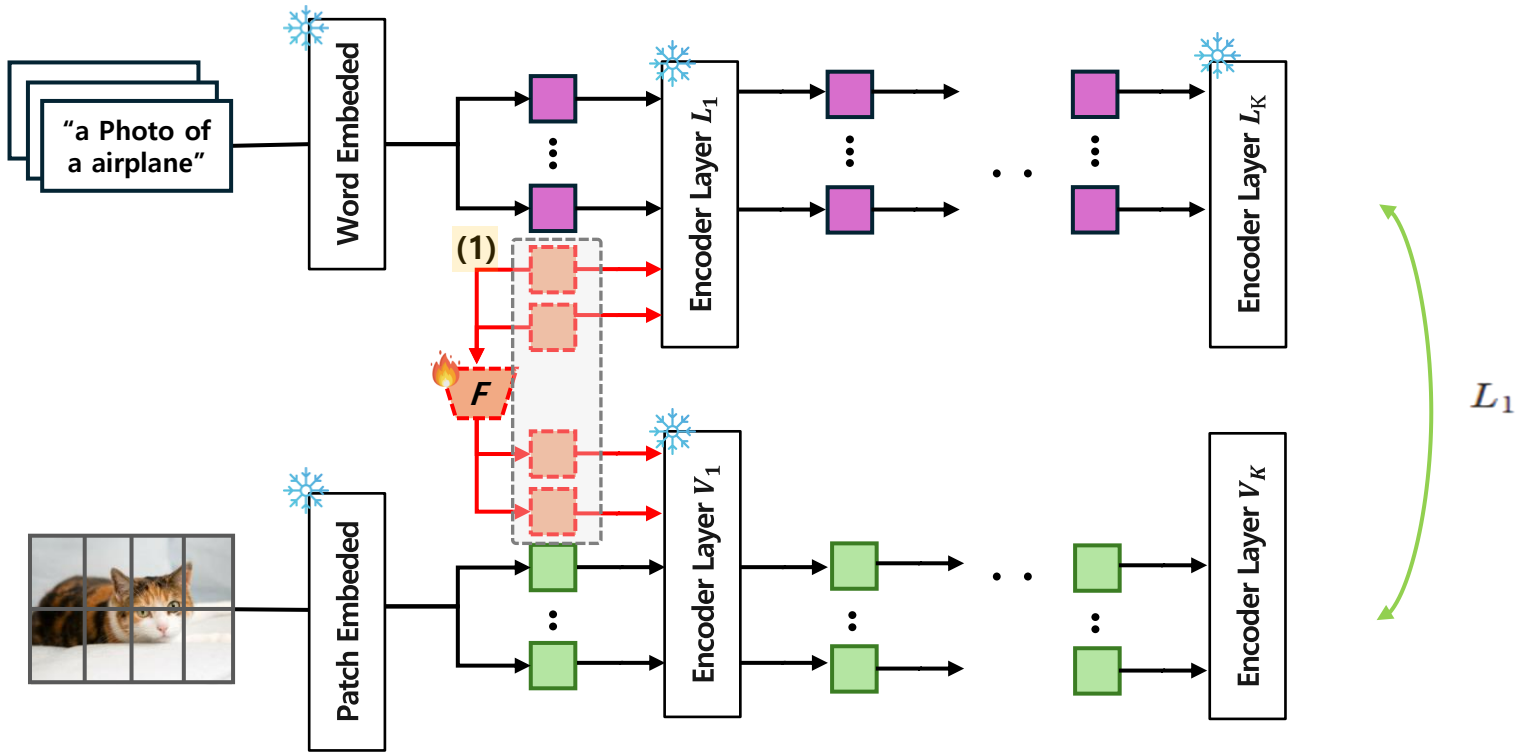
3. MaPLe

Multimodal Prompt Learning (MaPLe)

❖ MaPLe: Multimodal Prompt Learning [CVPR 2023]

(1) 제안 1 : Multimodal Prompting

- Text와 Image Encoder 모두에 learnable prompt를 삽입하여 두 modality를 공동으로 Task에 Adaptation
- 기존 Text-Only Prompt Learning을 Vision-Language Multimodal Prompting으로 확장



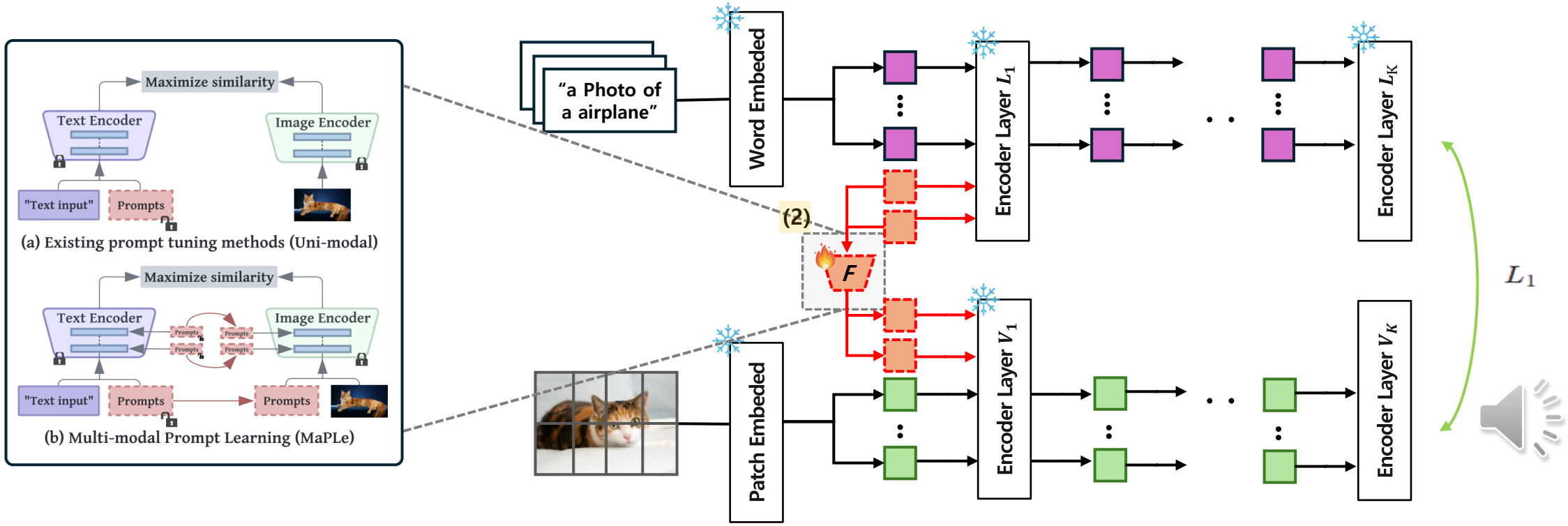
3. MaPLe

Multimodal Prompt Learning (MaPLe)

❖ MaPLe: Multi-modal Prompt Learning [CVPR 2023]

(2) 제안 2 : V-L coupling function

- 각 layer의 Language Prompt 를 Coupling Function F 에 입력하여 동일한 Layer 의 Visual Prompt 를 생성
- Visual Prompt를 독립적으로 학습하지 않고 Language Prompt로부터 projection하여 Vision-Language 정렬을 유지



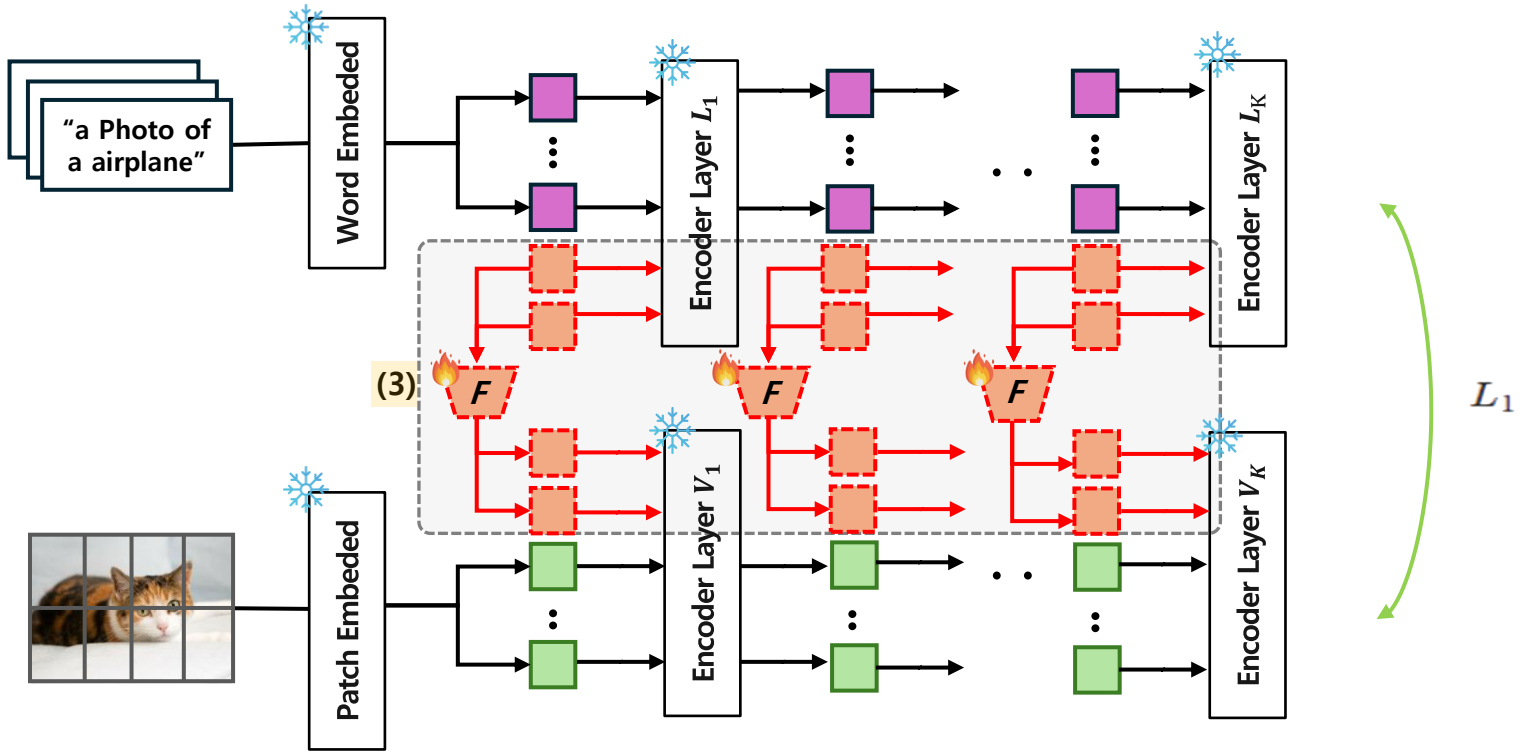
3. MaPLe

Multimodal Prompt Learning (MaPLe)

❖ MaPLe: Multi-modal Prompt Learning [CVPR 2023]

(3) 제안 3 : Deep Prompting

- 여러 Transformer layer에 서로 다른 Learnable Prompt를 삽입하여 인코더 깊이에 따라 Representation을 점진적으로 조정
- Layer별 Prompt가 서로 다른 수준의 Context 관계를 학습하여 **Stage-Wise Feature Representation 학습을 수행**



3. MaPLe

Multimodal Prompt Learning (MaPLe)

❖ Result

✓ **정성적 지표 (t-SNE)**

- CoCoOp은 Text prompt만 조정하므로 Image Representation은 CLIP과 동일한 반면, MaPLe은 Image Encoder까지 Task에 맞게 학습하였음
- DTD, EuroSAT, UCF101의 t-SNE에서 MaPLe이 CoCoOp보다 **Base와 Novel class 모두에서 더 명확한 class separation**을 형성
- Vision과 Language Representation을 함께 조정하는 **Multimodal Promp Learning 의 효과**를 시각적으로 확인

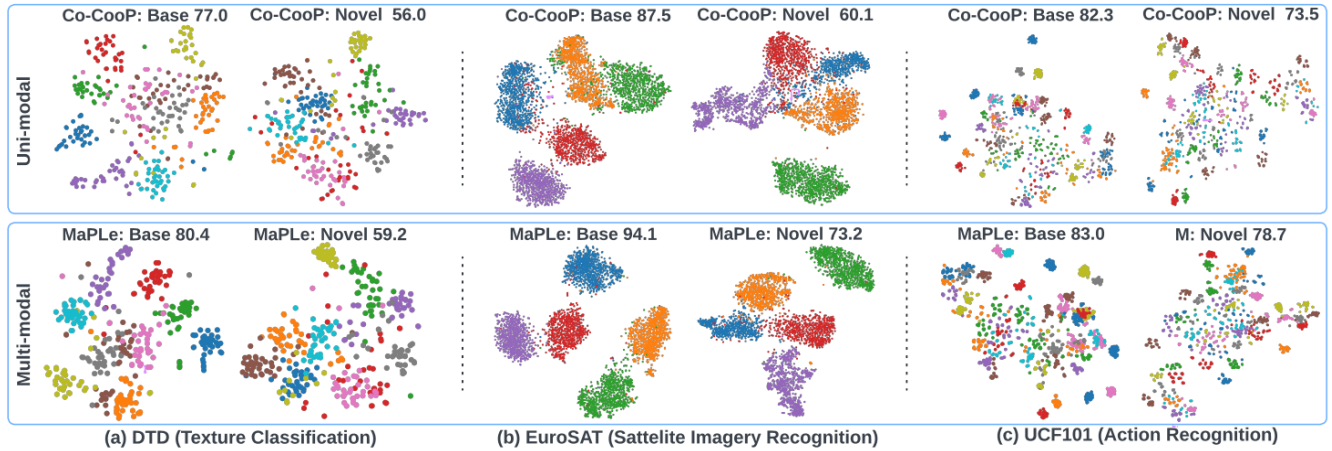


Figure 3. t-SNE plots of image embeddings in uni-modal prompting method Co-CoOp, and MaPLe on 3 diverse image recognition datasets. MaPLe shows better separability in both base and novel classes.

3. MaPLe

Multimodal Prompt Learning (MaPLe)

❖ Result

✓ **정량적 지표**

(1) Base-to-New Generalization (B2N)

- 11개 데이터셋 평균에서 CoCoOp 대비 **Base 80.47→82.28 / New 71.69→75.14 / HM 75.83→78.55**로 모두 향상
- Caltech101의 Base 성능을 제외한 모든 데이터셋에서 Base와 New 성능을 동시에 개선
- 학습 클래스 적응과 Novel class 일반화를 동시에 개선하여 기존의 Base–New trade-off를 완화

1) Base-to-New Generalization

(a) **Average over 11 datasets**

	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
MaPLe	82.28	**75.14**	**78.55**
	+1.81	+3.45	+2.72

(b) ImageNet.

	Base	Novel	HM
CLIP	72.43	68.14	70.22
CoOp	76.47	67.88	71.92
Co-CoOp	75.98	70.43	73.10
MaPLe	**76.66**	**70.54**	**73.47**
	+0.68	+0.11	+0.37

(c) Caltech101

	Base	Novel	HM
CLIP	96.84	94.00	95.40
CoOp	**98.00**	89.81	93.73
Co-CoOp	97.96	93.81	95.84
MaPLe	97.74	**94.36**	**96.02**
	-0.22	+0.55	+0.18

(d) OxfordPets

	Base	Novel	HM
CLIP	91.17	97.26	94.12
CoOp	93.67	95.29	94.47
Co-CoOp	95.20	97.69	96.43
MaPLe	**95.43**	**97.76**	**96.58**
	+0.23	+0.07	+0.15

(e) StanfordCars

	Base	Novel	HM
CLIP	63.37	**74.89**	68.65
CoOp	**78.12**	60.40	68.13
Co-CoOp	70.49	73.59	72.01
MaPLe	72.94	74.00	**73.47**
	+2.45	+0.41	+1.46

(f) Flowers102

	Base	Novel	HM
CLIP	72.08	**77.80**	74.83
CoOp	**97.60**	59.67	74.06
Co-CoOp	94.87	71.75	81.71
MaPLe	95.92	72.46	**82.56**
	+1.05	+0.71	+0.85

(g) Food101

	Base	Novel	HM
CLIP	90.10	91.22	90.66
CoOp	88.33	82.26	85.19
Co-CoOp	90.70	91.29	90.99
MaPLe	**90.71**	**92.05**	**91.38**
	+0.01	+0.76	+0.39

(h) FGVCAircraft

	Base	Novel	HM
CLIP	27.19	**36.29**	31.09
CoOp	**40.44**	22.30	28.75
Co-CoOp	33.41	23.71	27.74
MaPLe	37.44	35.61	**36.50**
	+4.03	+11.90	+8.76

(i) SUN397

	Base	Novel	HM
CLIP	69.36	75.35	72.23
CoOp	80.60	65.89	72.51
Co-CoOp	79.74	76.86	78.27
MaPLe	**80.82**	**78.70**	**79.75**
	+1.08	+1.84	+1.48

(j) DTD

	Base	Novel	HM
CLIP	53.24	**59.90**	56.37
CoOp	79.44	41.18	54.24
Co-CoOp	77.01	56.00	64.85
MaPLe	**80.36**	59.18	**68.16**
	+3.35	+3.18	+3.31

(k) EuroSAT

	Base	Novel	HM
CLIP	56.48	64.05	60.03
CoOp	92.19	54.74	68.69
Co-CoOp	87.49	60.04	71.21
MaPLe	**94.07**	**73.23**	**82.35**
	+6.58	+13.19	+11.14

(l) UCF101

	Base	Novel	HM
CLIP	70.53	77.50	73.85
CoOp	**84.69**	56.05	67.46
Co-CoOp	82.33	73.45	77.64
MaPLe	83.00	**78.66**	**80.77**
	+0.67	+5.21	+3.13

3. MaPLe

Multimodal Prompt Learning (MaPLe)

❖ Result

✓ **정량적 지표**

(2) Cross-Dataset Evaluation (CD)

- ImageNet 성능은 기존 방법과 유사하지만, 10개 target dataset 평균은 CoCoOp 65.74→MaPLe 66.30으로 향상
- DTD, EuroSAT, FGVCAircraft 등 ImageNet과 분포 차이가 큰 데이터셋에서 비교적 큰 개선을 기록

(3) Domain Generalization (DG)

- ImageNetV2에서는 CoCoOp과 동일한 성능을 보이고, ImageNet-Sketch, ImageNet-A, ImageNet-R에서는 모두 향상
- Multimodal prompting을 통해 스타일과 표현 방식이 달라지는 Domain Shift에 대한 강건성 개선

2) Cross-Dataset Evaluation

	Source	Target									
	ImageNet	Caltech101	OxfordPets	StanfordCars	Flowers102	Food101	Aircraft	SUN397	DTD	EuroSAT	UCF101	Average
CoOp	**71.51**	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	46.39	66.55	63.88
Co-CoOp	71.02	**94.43**	90.14	65.32	71.88	86.06	22.94	**67.36**	45.73	45.37	68.21	65.74
MaPLe	70.72	93.53	**90.49**	**65.57**	**72.23**	**86.20**	**24.74**	67.01	**46.49**	**48.06**	**68.69**	**66.30**

3) Domain Generalization

	Source	Target			
	ImageNet	ImageNetV2	ImageNet-S	ImageNet-A	ImageNet-R
CLIP	66.73	60.83	46.15	47.77	73.96
CoOp	**71.51**	**64.20**	47.99	49.71	75.21
Co-CoOp	71.02	64.07	48.75	50.63	76.18
MaPLe	70.72	64.07	**49.15**	**50.90**	**76.98**

3. MaPLe

Multimodal Prompt Learning (MaPLe)

❖ Ablation Study

(1) Multimodal Prompting 추가 실험

- Deep vision prompting보다 Deep language prompting의 성능이 높은 것으로 보아, CLIP Adaptatation에서 언어 브랜치의 Prompt가 더 큰 기여를 하는 것을 확인
- 두 Modality에 Prompt를 단순히 독립적으로 삽입하는 것보다 **상호 연관된 학습이 중요함**을 확인

(2) V-L coupling function 추가 실험

- Language Prompt에서 Visual Prompt로 projection하는 $P \rightarrow \tilde{P}$ 구조가 반대 방향보다 우수
- Language → Vision Coupling 방식이 modality 간 정보 손실을 줄이고 정렬 유지에 더 효과적임을 확인

(1) Multi-modal Prompting

Method	Base Acc.	Novel Acc.	HM	GFLOPS
1: MaPLe shallow ($J = 1$)	80.10	73.52	76.67	167.1
2: Deep vision prompting	80.24	73.43	76.68	18.0
3: Deep language prompting	81.72	73.81	77.56	166.8
4: Independent V-L prompting	82.15	74.07	77.90	167.0
5: MaPLe (Ours)	**82.28**	**75.14**	**78.55**	167.0

Table 1. Comparison of MaPLe with different prompting designs in base-to-novel generalization. Results are averaged over 11 datasets. HM refers to harmonic mean.

(2) V-L coupling function

Prompt Proj.	Base	Novel	HM
$\tilde{P} \rightarrow P$	81.37	73.25	77.10
$P \rightarrow P$	**82.28**	**75.14**	**78.55**

Table 9. Projecting from P to $\tilde{P}$ provides the best results.

3. MaPLe

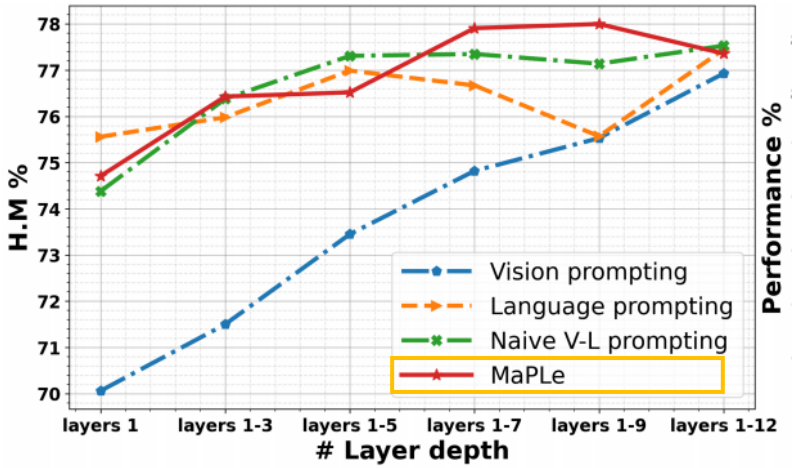
Multimodal Prompt Learning (MaPLe)

❖ Ablation Study

(3) Deep Prompting 추가 실험

- Prompt를 더 많은 Transformer layer에 삽입할수록 초기에는 성능이 점진적으로 향상
- MaPLe은 **layer depth 9에서 가장 높은 HM 성능**을 기록하며, Shallow Prompting보다 Deep Prompting이 효과적임을 확인
- 모든 layer에 prompt를 삽입할 경우 성능이 소폭 감소하여, 지나치게 깊은 prompting은 과적합을 유발할 가능성이 존재

(3) Deep Prompting



3. MaPLe

Multimodal Prompt Learning (MaPLe)

❖ Model Efficiency

(1) Params

- 공통 coupling function을 사용하는 MaPLe†는 파라미터를 **3.55M→0.41M**으로 약 9배 줄여도 HM 78.11을 유지

  → 따라서, 성능 향상이 **단순한 파라미터 증가가 아니라 Multimodal Prompting 구조에서 기인함**을 확인

(2) 추론 효율 (FPS)

- CoCoOp은 이미지별 Text Feature 계산으로 batch size가 증가해도 FPS가 거의 향상되지 않음
- 그러나, MaPLe은 **class 의 Text Feature를 효율적으로 처리하여 batch size 증가에 따라 FPS가 크게 향상**

Method	Params	Params % CLIP	FPS (with BS) 1	4	100	HM
CoOp	2048	0.002	13.8	55.3	1353.0	71.66
CoCoOp	35360	0.03	64.6	114.7	15.1	75.83
Independent V-L	31488	0.02	62.5	239.4	1383.8	77.90
MaPLe	3.55 M	2.85	60.2	239.0	1365.1	78.55
MaPLe†	0.41 M	0.33	60.2	238.0	1365.0	78.11

Table 6. Comparison of computational complexity among different prompting methods. MaPLe† is a MaPLe version which utilizes a common V-L coupling function for all layers.

3. MaPLe

Multimodal Prompt Learning (MaPLe)

❖ Prompt Learning 의 발전과정

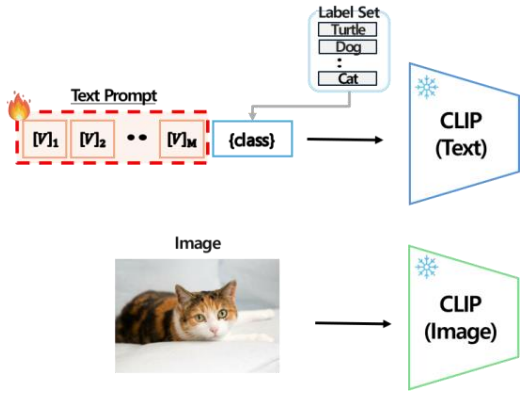
Text Prompt | **Multimodal Prompt** | **Multimodal Representation**

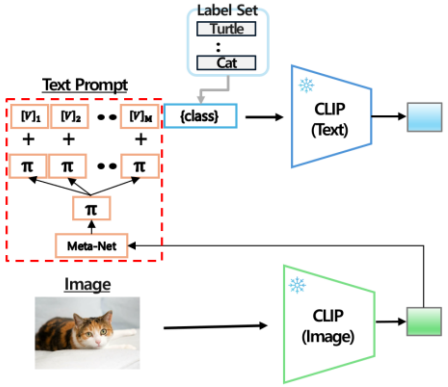
CoOp [IJCV 2022] → **CoCoOp** [CVPR 2022] → **MaPLe** [CVPR 2023] → **MMRL** [CVPR 2025]

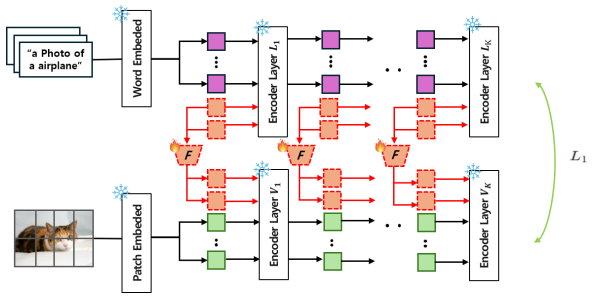
연속적인 Context vector 제안

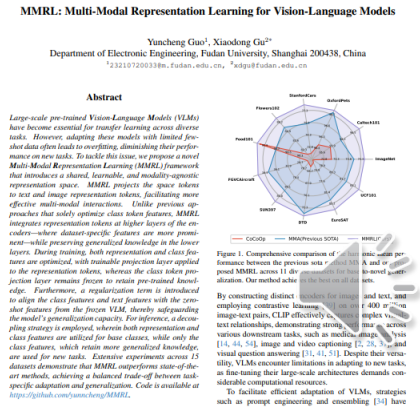
조건화된 Prompt Learning 제안

{Text + Visual}
Prompt Learning 제안








MMRL: Multi-Modal Representation Learning for Vision-Language Models

Yuncheng Guo[1], Xiaodong Gu[2*]
Department of Electronic Engineering, Fudan University, Shanghai 200438, China
[1]23210720033@m.fudan.edu.cn, [2*]xdgu@fudan.edu.cn

Abstract

4. MMRL

Multi-Modal Representation Learning (MMRL)

❖ Prompt Learning 의 발전과정

Text Prompt | **Multimodal Prompt** | **Multimodal Representation**

한계 :
V-L 연결에서 Text 중심의
Prompt Learning

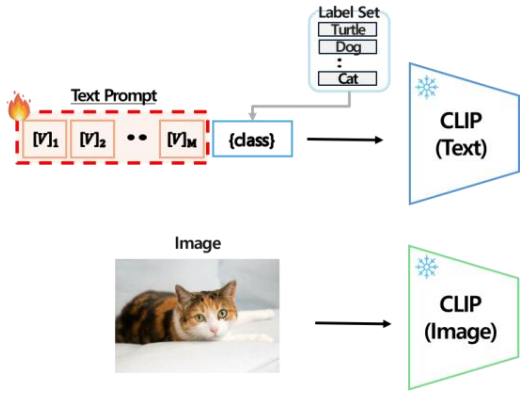
CoOp [IJCV 2022] → **CoCoOp** [CVPR 2022] → **MaPLe** [CVPR 2023] → **MMRL** [CVPR 2025]

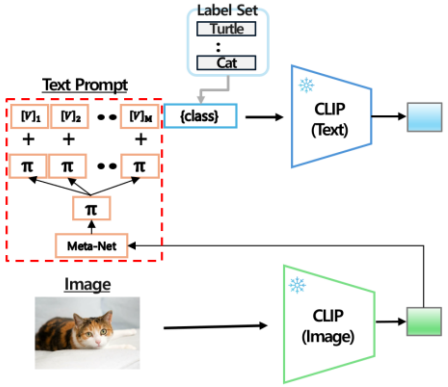
연속적인 Context vector 제안

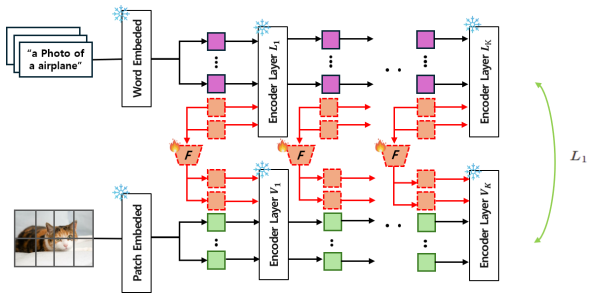
조건화된 Prompt Learning 제안

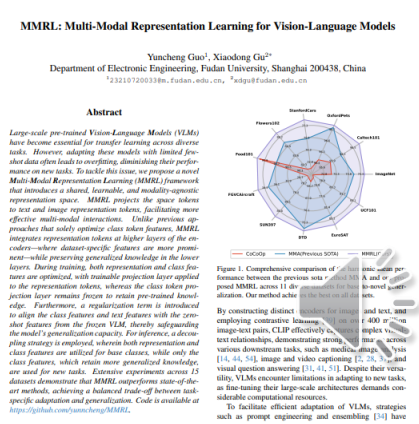
{Text + Visual}
Prompt Learning 제안

Multimodel Representation Learning








MMRL: Multi-Modal Representation Learning for Vision-Language Models

Yuncheng Guo[1], Xiaodong Gu[2*]
Department of Electronic Engineering, Fudan University, Shanghai 200438, China
[1]23210720033@m.fudan.edu.cn, [2*]xdgu@fudan.edu.cn

Abstract

Large-scale pre-trained Vision-Language Models (VLMs) have become essential for transfer learning across diverse tasks. However, adapting these models with limited few-shot data often leads to overfitting, diminishing their performance on new tasks. To tackle this issue, we propose a novel Multi-Modal Representation Learning (MMRL) framework that introduces a shared, learnable, and modality-agnostic representation space. MMRL projects the space tokens to text and image representation tokens, facilitating more effective multi-modal interactions. Unlike previous approaches that solely optimize class token features, MMRL integrates representation tokens at higher layers of the encoders—where dataset-specific features are more prominent—while preserving generalized knowledge in the lower layers. During training, both representation and class features are optimized, with trainable projection layer applied to the representation tokens, whereas the class token projection layer remains frozen to retain pre-trained knowledge. Furthermore, a regularization term is introduced to align the class features and text features with the zero-shot features from the frozen VLM, thereby safeguarding the model's generalization capacity. For inference, a decoupling strategy is employed, wherein both representation and class features are utilized for base classes, while only the class features, which retain more generalized knowledge, are used for new tasks. Extensive experiments across 15 datasets demonstrate that MMRL outperforms state-of-the-art methods, achieving a balanced trade-off between task-specific adaptation and generalization. Code is available at https://github.com/yunncheng/MMRL.

Figure 1. Comprehensive comparison of ... with state-of-the-art methods on performance between the previous sota ... and our proposed MMRL across 11 ... for base-to-novel generalization. Our method achie... on all datasets.

By constructing distinct ... encoders for image ... and text, and employing contrastive learning ... on over 400 million image-text pairs, CLIP effectively ... complex ... text relationships, demonstrating strong ... across various downstream tasks, such as medical ... analysis [14, 44, 54], image and video captioning [2, 28, 3, 4], and visual question answering [31, 41, 51]. Despite their versatility, VLMs encounter limitations in adapting to new tasks, as fine-tuning their large-scale architectures demands considerable computational resources.

To facilitate efficient adaptation of VLMs, strategies such as prompt engineering and ensembling [34] have

4. MMRL

Multi-Modal Representation Learning (MMRL)

❖ MaPLe 은 어떤 한계가 있었을까?

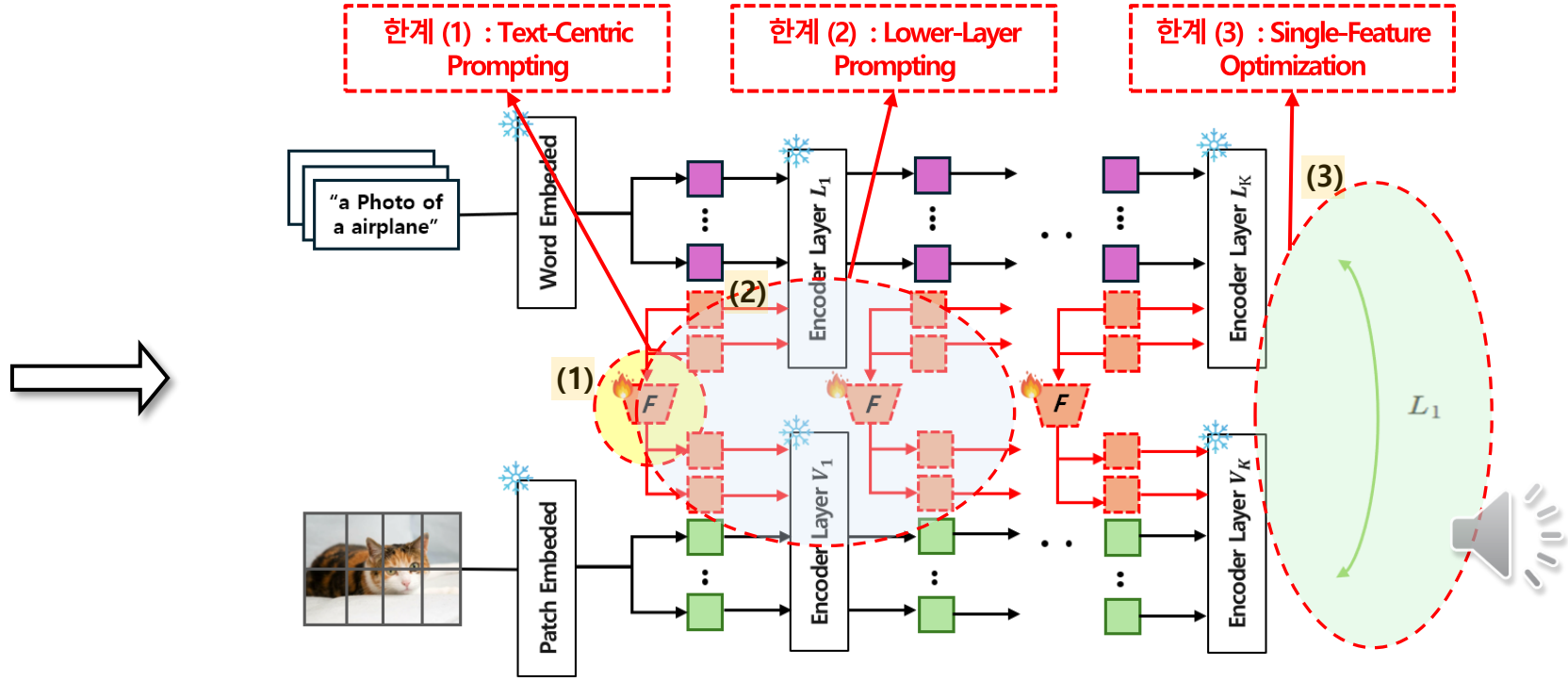
(1) **한계 1** : Text-Centric Prompting

(2) **한계 2 :** Lower-Layer Prompting

(3) **한계 3 :** Single-Feature Optimization

→ **MMRL** [4] **: 상위 Layer의 Shared representatio을 기반으로 Adaptation과 Generalization을 분리하는 구조 제안**

MaPLe
[CVPR 2023]



4. MMRL

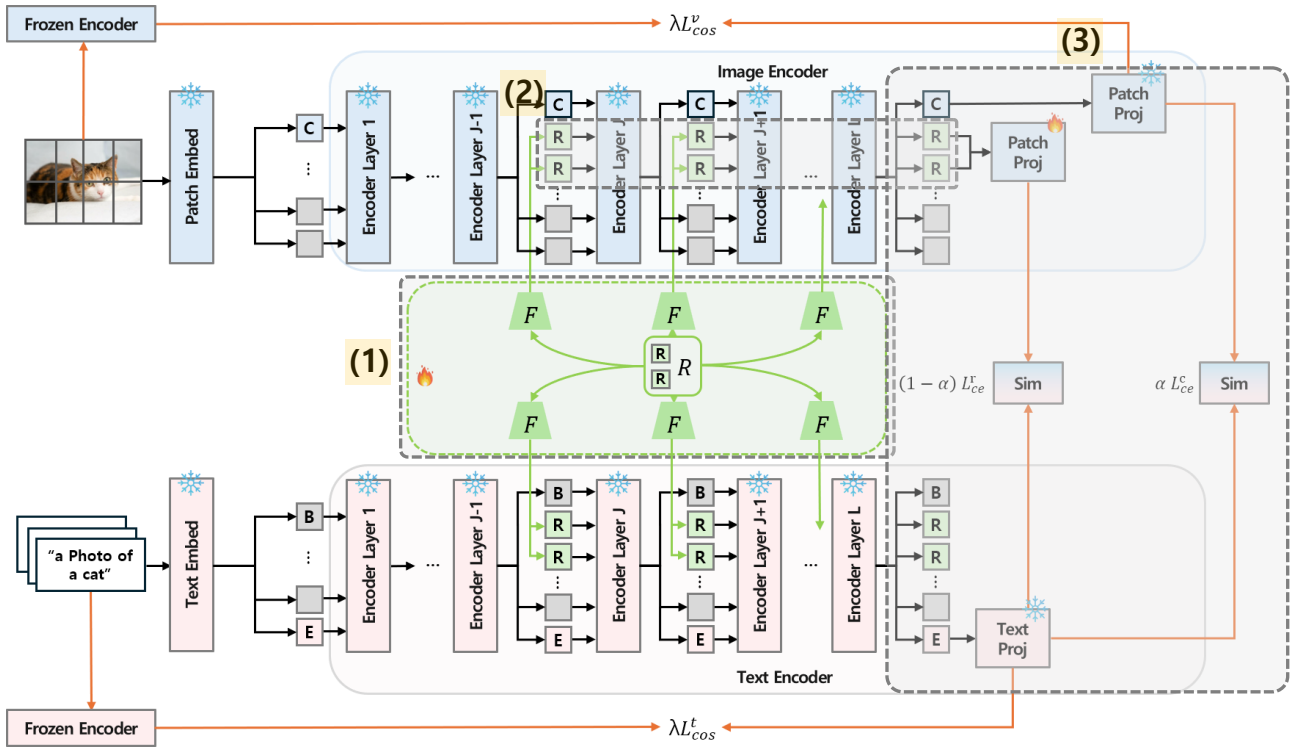
Multi-Modal Representation Learning (MMRL)

❖ MMRL: Multi-Modal Representation Learning for Vision-Language Models [CVPR 2025]

(1) 제안 1 : Shared Representation Space

(2) 제안 2 : Higher-Layer Representation Learning

(3) 제안 3 : Decoupling Strategy



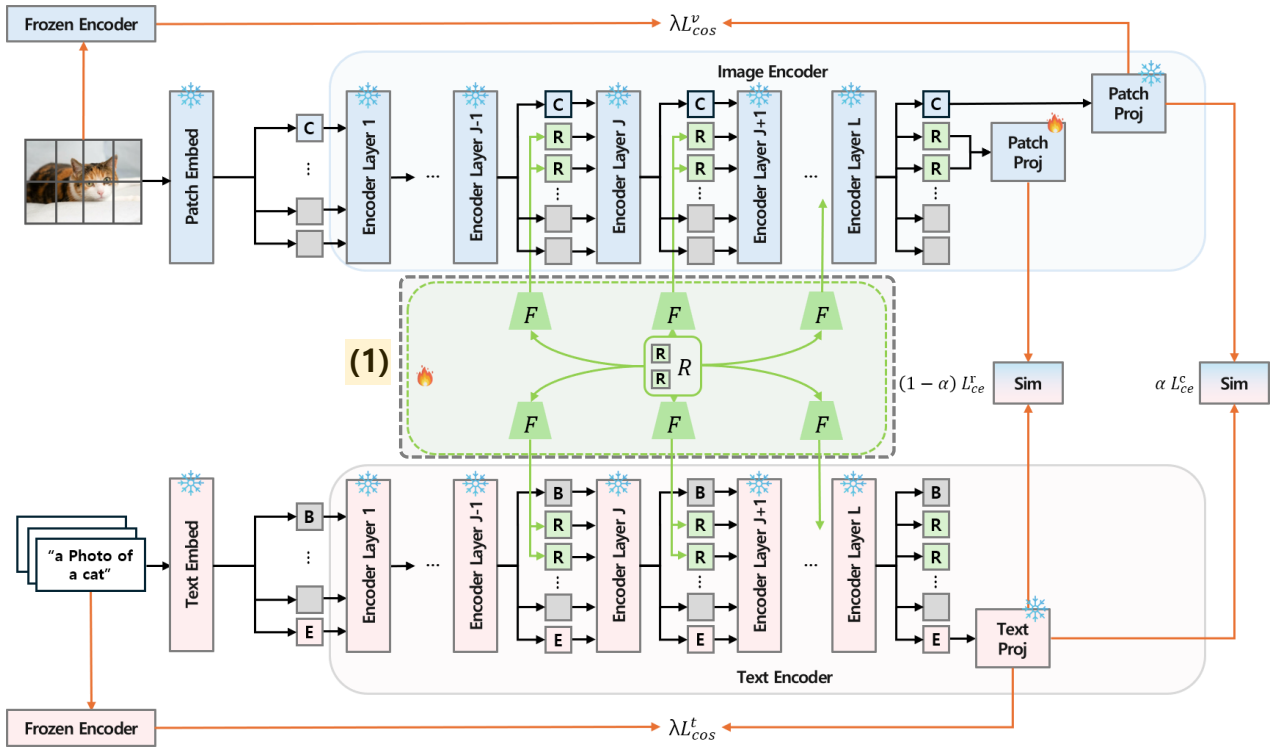
4. MMRL

Multi-Modal Representation Learning (MMRL)

❖ MMRL: Multi-Modal Representation Learning for Vision-Language Models [CVPR 2025]

(1) 제안 1 : Shared Representation Space

- Modality-agnostic shared representation $\mathcal{R}$에서 Visual / Textual Representation token을 각각 생성
- Text 중심의 단방향 projection을 넘어 두 modality가 동일한 Representation Space를 공동 학습



4. MMRL

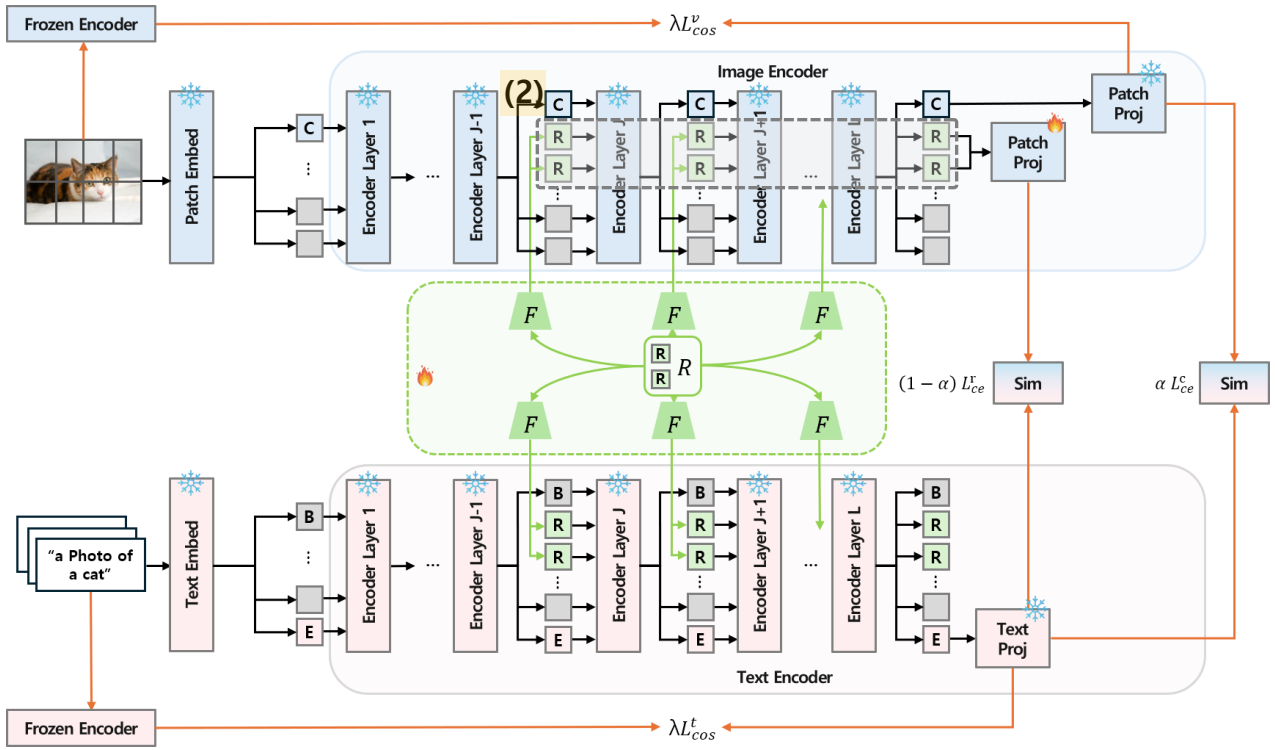
Multi-Modal Representation Learning (MMRL)

❖ MMRL: Multi-Modal Representation Learning for Vision-Language Models [CVPR 2025]

(2) 제안 2 : Higher-Layer Representation Learning

- Lower layer는 pretrained CLIP의 일반화된 특징을 보존하고, Representation token은 J번째 이후 higher layer에만 삽입

→ Task-specific adaptation을 encoder 후반부에 제한하여 few-shot overfitting을 완화



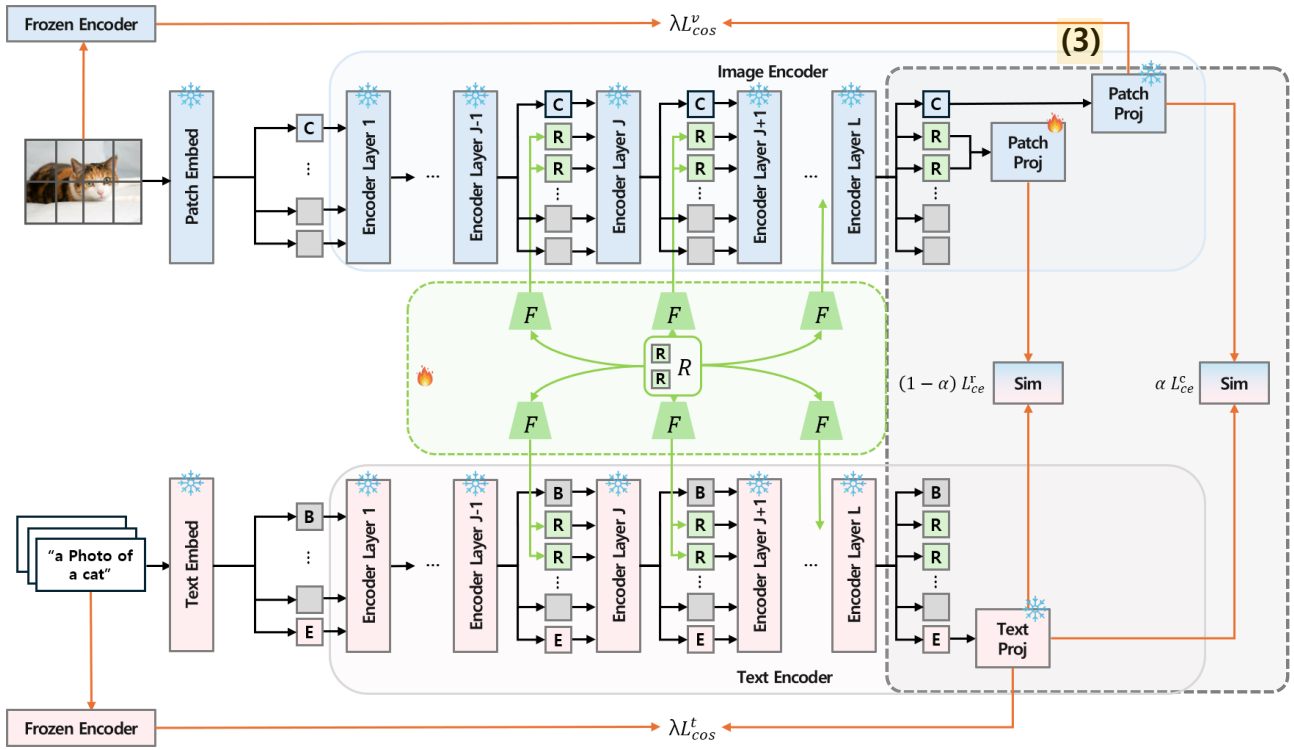
4. MMRL

Multi-Modal Representation Learning (MMRL)

❖ MMRL: Multi-Modal Representation Learning for Vision-Language Models [CVPR 2025]

(3) 제안 3 : **Decoupling Strategy**

- **Class feature:** 기존 CLS/EOT token과 frozen projection을 사용하여 pretrained CLIP의 generalization knowledge를 유지
- **Representation feature:** 새롭게 학습한 representation token과 trainable projection을 통해 task-specific knowledge를 학습



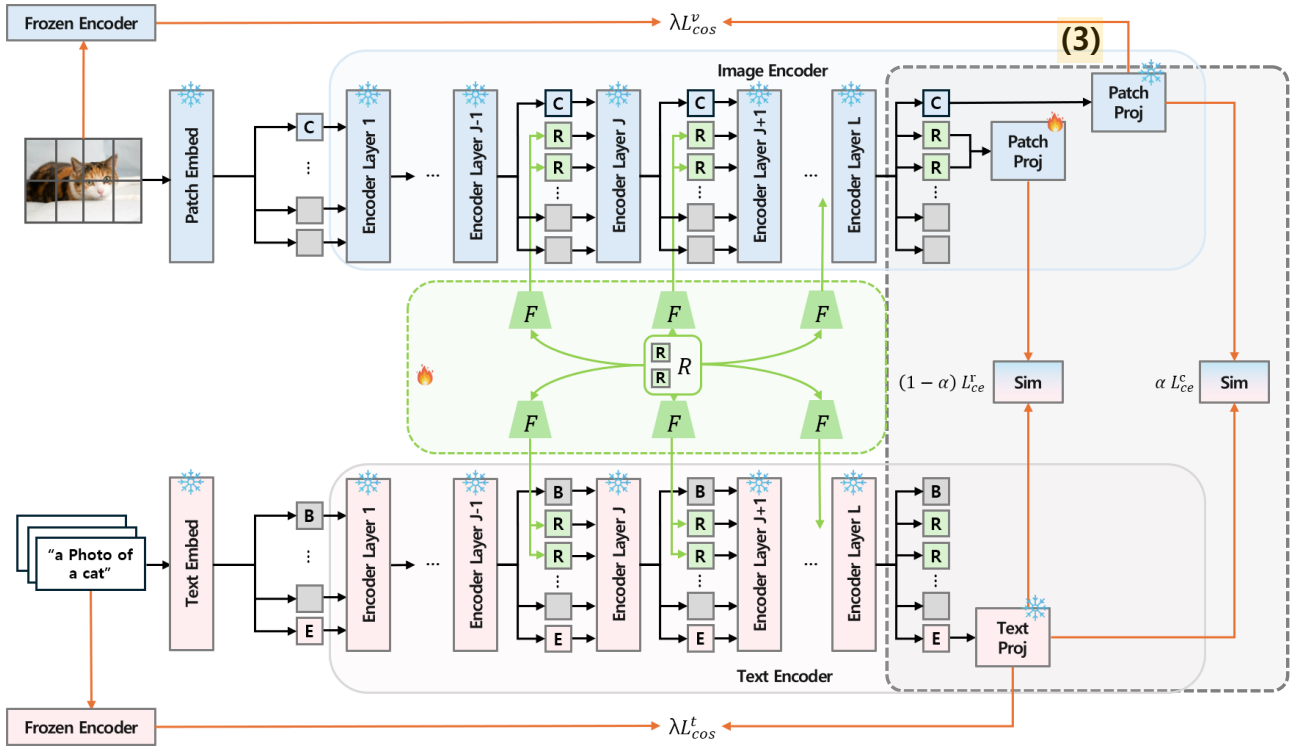
4. MMRL

Multi-Modal Representation Learning (MMRL)

❖ MMRL: Multi-Modal Representation Learning for Vision-Language Models [CVPR 2025]

(3) 제안 3 : Decoupling Strategy (Inference)

- Base class에서는 class feature와 representation feature를 결합하여 adaptation과 generalization을 함께 활용
- Novel class 및 새로운 task에서는 class feature만 사용하여 학습 데이터에 치중되는 overfitting 영향 차단



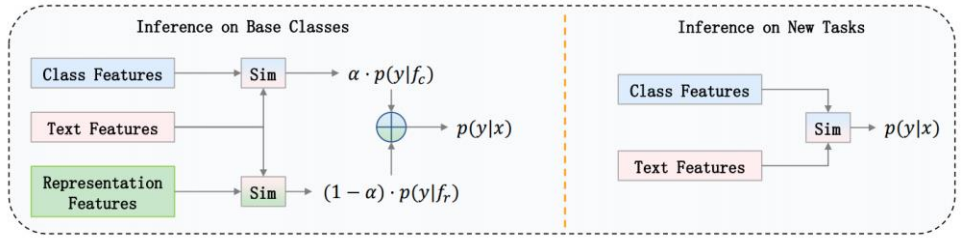


Figure 3. MMRL inference process, where different tasks utilize distinct features.

4. MMRL

Multi-Modal Representation Learning (MMRL)

❖ Result

(1) Base-to-New Generalization (B2N)

- 11개 데이터셋 평균에서 **Base 85.68 / Novel 77.16 / HM 81.20** 3가지 모두 최고 성능을 기록
- 일부 데이터셋에서 최고 Novel accuracy는 아니지만, Base와 Novel의 균형 측면에서 가장 높은 HM을 기록

1) Base-to-New Generalization

Method	Average			ImageNet			Caltech101			OxfordPets		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CLIP (ICML2021)	69.34	74.22	71.70	72.43	68.14	70.22	96.84	94.00	95.40	91.17	97.26	94.12
CoOp (IJCV2022)	82.69	63.22	71.66	76.47	67.88	71.92	98.00	89.81	93.73	93.67	95.29	94.47
CoOpOp (CVPR2022)	80.47	71.69	75.83	75.98	70.43	73.10	97.96	93.81	95.84	95.20	97.69	96.43
ProDA (CVPR2022)	81.56	72.30	76.65	75.40	70.23	72.72	98.27	93.23	95.68	95.43	97.83	96.62
KgCoOp (CVPR2023)	80.73	73.60	77.00	75.83	69.96	72.78	97.72	94.39	96.03	94.65	97.76	96.18
MaPLe (CVPR2023)	82.28	75.14	78.55	76.66	70.54	73.47	97.74	94.36	96.02	95.43	97.76	96.58
PromptSRC (ICCV2023)	84.26	76.10	79.97	77.60	70.73	74.01	98.10	94.03	96.02	95.33	97.30	96.30
ProVP (IJCV2024)	85.20	73.22	78.76	75.82	69.21	72.36	98.92	94.21	96.51	95.87	97.65	**96.75**
MetaPrompt (TIP2024)	83.65	75.48	79.09	77.52	70.83	74.02	98.13	94.58	96.32	95.53	97.00	96.26
TCP (CVPR2024)	84.13	75.36	79.51	77.27	69.87	73.38	98.23	**94.67**	96.42	94.67	97.20	95.92
MMA (CVPR2024)	83.20	76.80	79.87	77.31	71.00	74.02	98.40	94.00	96.15	95.40	**98.07**	96.72
MMRL (Ours)	**85.68**	**77.16**	**81.20**	**77.90**	**71.30**	**74.45**	**98.97**	94.50	**96.68**	**95.90**	97.60	96.74

Method	StanfordCars			Flowers102			Food101			FGVCAircraft		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CLIP (ICML2021)	63.37	74.89	68.65	72.08	77.80	74.83	90.10	91.22	90.66	27.19	36.29	31.09
CoOp (IJCV2022)	78.12	60.40	68.13	97.60	59.67	74.06	88.33	82.26	85.19	40.44	22.30	28.75
CoOpOp (CVPR2022)	70.49	73.59	72.01	94.87	71.75	81.71	90.70	91.29	90.99	33.41	23.71	27.74
ProDA (CVPR2022)	74.70	71.20	72.91	97.70	68.68	80.66	90.30	88.57	89.43	36.90	34.13	35.46
KgCoOp (CVPR2023)	71.76	75.04	73.36	95.00	74.73	83.65	90.50	91.70	91.09	36.21	33.55	34.83
MaPLe (CVPR2023)	72.94	74.00	73.47	95.92	72.46	82.56	90.71	**92.05**	**91.38**	37.44	35.61	36.50
PromptSRC (ICCV2023)	78.27	74.97	76.58	98.07	76.50	85.95	90.67	91.53	91.10	42.73	**37.87**	40.15
ProVP (IJCV2024)	80.43	67.96	73.67	98.42	72.06	83.20	90.32	90.91	90.61	**47.08**	29.87	36.55
MetaPrompt (TIP2024)	76.34	75.01	75.48	97.66	74.49	84.52	**90.74**	91.85	91.29	40.14	36.51	38.24
TCP (CVPR2024)	80.80	74.13	77.32	97.73	75.57	85.23	90.57	91.37	90.97	41.97	34.43	37.83
MMA (CVPR2024)	78.50	73.10	75.70	97.77	75.93	85.48	90.13	91.30	90.71	40.57	36.33	38.33
MMRL (Ours)	**81.30**	**75.07**	**78.06**	**98.97**	**77.27**	**86.78**	90.57	91.50	91.03	46.30	37.03	**41.15**

Method	SUN397			DTD			EuroSAT			UCF101		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CLIP (ICML2021)	69.36	75.35	72.23	53.24	59.90	56.37	56.48	64.05	60.03	70.53	77.50	73.85
CoOp (IJCV2022)	80.60	65.89	72.51	79.44	41.18	54.24	92.19	54.74	68.69	84.69	56.05	67.46
CoOpOp (CVPR2022)	79.74	76.86	78.27	77.01	56.00	64.85	87.49	60.04	71.21	82.33	73.45	77.64
ProDA (CVPR2022)	78.67	76.93	77.79	80.67	56.48	66.44	83.90	66.00	73.88	85.23	71.97	78.04
KgCoOp (CVPR2023)	80.29	76.53	78.36	77.55	54.99	64.35	85.64	64.34	73.48	82.89	76.67	79.65
MaPLe (CVPR2023)	80.82	78.70	79.75	80.36	59.18	68.16	94.07	73.23	82.35	83.00	78.66	80.77
PromptSRC (ICCV2023)	82.67	78.47	80.52	83.37	62.97	71.75	92.90	73.90	82.32	87.10	78.80	82.74
ProVP (IJCV2024)	80.67	76.11	78.32	83.95	59.06	69.34	**97.12**	72.91	83.29	**88.56**	75.55	81.54
MetaPrompt (TIP2024)	82.26	79.04	80.62	83.10	58.05	68.35	93.53	75.21	83.38	85.33	77.72	81.35
TCP (CVPR2024)	82.63	78.20	80.35	82.77	58.07	68.25	91.63	74.73	82.32	87.13	**80.77**	83.83
MMA (CVPR2024)	82.27	78.57	80.38	83.20	**65.63**	73.38	85.46	**82.34**	83.87	86.23	80.03	82.20
MMRL (Ours)	**83.20**	**79.30**	**81.20**	**85.67**	65.00	**73.82**	95.60	80.17	**87.21**	88.10	80.07	**83.89**

4. MMRL

Multi-Modal Representation Learning (MMRL)

❖ Result

(2) Cross-Dataset Evaluation (CD)

- 10개 Target Dataset 평균에서 67.25%를 기록하여 기존 방법보다 높은 Cross-Dataset transfer 성능 달성
- 다양한 Target Domain 전체에 대한 평균 일반화 성능에서 강점

(3) Domain Generalization (DG)

- 4개 Target Domain 중 ImageNetV2와 ImageNet-A에서 최고 성능 기록
- 다양한 Domain Shift에서도 사전학습된 general knowledge를 안정적으로 유지

2) Cross-Dataset Evaluation

	Source	Target										
	ImageNet	Average	Caltech101	OxfordPets	StanfordCars	Flowers101	Food101	FGVCAircraft	SUN397	DTD	EuroSAT	UCF101
CoOp (IJCV2022)	71.51	63.88	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	46.39	66.55
CoOpOp (CVPR2022)	71.02	65.74	94.43	90.14	65.32	71.88	86.06	22.94	67.36	45.73	45.37	68.21
MaPLe (CVPR2023)	70.72	66.30	93.53	90.49	65.57	72.23	86.20	24.74	67.01	46.49	48.06	68.69
PromptSRC (ICCV2023)	71.27	65.81	93.60	90.25	65.70	70.25	86.15	23.90	67.10	**46.87**	45.50	**68.75**
TCP (CVPR2024)	71.40	66.29	93.97	91.25	64.69	71.21	**86.69**	23.45	67.15	44.35	51.45	68.73
MMA (CVPR2024)	71.00	66.61	93.80	90.30	**66.13**	72.07	86.12	25.33	**68.17**	46.57	49.24	68.32
MMRL (Ours)	**72.03**	**67.25**	**94.67**	**91.43**	66.10	**72.77**	86.40	**26.30**	67.57	45.90	**53.10**	68.27

3) Domain Generalization

	Source	Target			
	ImageNet	-V2	-S	-A	-R
CLIP (ICML2021)	66.73	60.83	46.15	47.77	73.96
CoOp (IJCV2022)	71.51	64.20	47.99	49.71	75.21
CoOpOp (CVPR2022)	71.02	64.07	48.75	50.63	76.18
MaPLe (CVPR2023)	70.72	64.07	49.15	50.90	76.98
PromptSRC (ICCV2023)	71.27	64.35	**49.55**	50.90	**77.80**
MMA (CVPR2024)	71.00	64.33	49.13	51.12	77.32
MMRL (Ours)	**72.03**	**64.47**	49.17	**51.20**	77.53

4. MMRL

Multi-Modal Representation Learning (MMRL)

❖ Ablation Study

(1) Shared Representation Space 추가 실험

- **w/o L, w/o V :** 한쪽 modality의 representation token을 제거하면 HM 성능이 하락하며, 특히 Visual Representation을 제거할 때 성능 감소가 큰 것으로 보아 Visual branch가 Downstream Adaptation에 크게 기여함을 확인
- **w/o RS** : Shared space를 제거할 경우 Novel accuracy가 감소하여 modality-agnostic representation이 일반화에 기여함을 확인

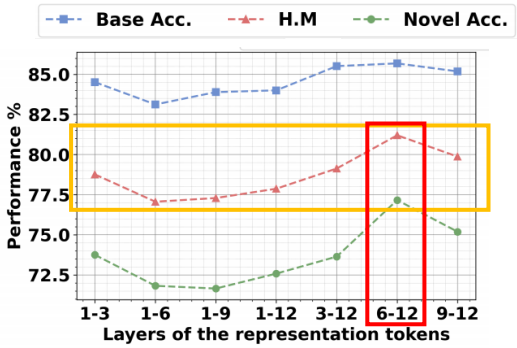
(2) Higher-Layer Representation Learning 추가 실험

- Lower layer 삽입은 pretrained knowledge를 손상시켜 성능이 저하
- **6–12 layer에서 가장 높은 HM**을 기록하여 상위 계층에서의 adaptation 효과를 확인

(1) Shared Representation Space & (3) Decoupling Strategy

Variants	Base	Novel	HM
w/o L	85.05	75.65	80.08
w/o V	82.83	75.03	78.74
w/o DS_1	83.59	77.16	80.25
w/o DS_2	85.68	73.80	79.30
w/o RS	**85.79**	75.55	80.34
MMRL†	85.60	76.02	80.55
MMRL	85.68	**77.16**	**81.20**

(2) Higher-Layer Representation Learning



4. MMRL

Multi-Modal Representation Learning (MMRL)

❖ Ablation Study

(3) Decoupling Strategy 추가 실험

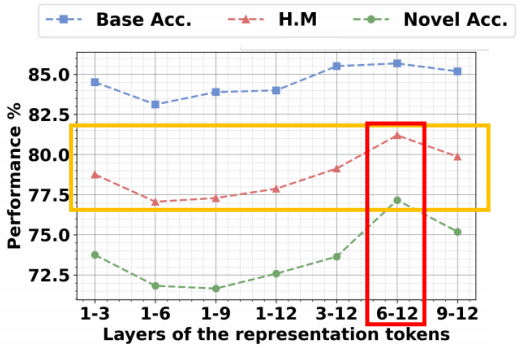
- **w/o DS_1 :** Base class에서 Representation feature를 제거하면 Base accuracy가 하락하여 task-specific adaptation 효과를 확인
- **w/o DS_2:** Novel class에도 representation feature를 사용하면 Novel accuracy가 **77.16→73.80**으로 크게 하락

⇒ **Base와 Novel에 서로 다른 feature를 사용하는 Decoupled Strategy 의 필요성**을 검증

(1) Shared Representation Space & (3) Decoupling Strategy

Variants	Base	Novel	HM
w/o L	85.05	75.65	80.08
w/o V	82.83	75.03	78.74
w/o DS_1	83.59	77.16	80.25
w/o DS_2	85.68	73.80	79.30
w/o RS	**85.79**	75.55	80.34
MMRL†	85.60	76.02	80.55
MMRL	85.68	**77.16**	**81.20**

(2) Higher-Layer Representation Learning



4. MMRL

Multi-Modal Representation Learning (MMRL)

❖ Model Efficiency

(1) Params

- 기본 MMRL은 **4.992M parameters**로 가장 작은 모델은 아니지만, HM 81.20으로 가장 높은 성능을 기록
- Representation dimension을 축소한 MMRL*은 parameter를 0.689M까지 줄이면서도 HM 80.84를 유지

(2) 학습 시간

- MMA보다는 느리지만, 높은 성능을 고려하면 학습 시간과 정확도 사이에서 우수한 균형을 제공

(3) 추론 효율 (FPS)

- BS=100에서 762.4 FPS로 MaPLe, PromptSRC 보다는 느리지만, 높은 성능과 빠른 학습 속도를 가짐

Method	Modality	Params (learnable)	Train time (ms/image)	Train time (minute/all)	FPS (100 BS)	HM
MaPLe	V-L	3.555M	39.5	26.4	1757.6	78.55
PromptSRC	V,L	0.046M	40.0	106.8	1764.2	79.97
ProVP	V	0.147M	4.4	107.2	928.9	78.76
MetaPrompt	V,L	0.031M	30.7	32.8	659.8	79.09
TCP	L	0.332M	5.3	17.7	950.6	79.51
MMA	V-L	0.675M	2.2	1.5	688.5	79.87
MMRL	V-L	4.992M	5.3	3.6	762.4	81.20
MMRL*	V-L	0.689M	5.3	3.6	767.8	80.84

4. MMRL

Multi-Modal Representation Learning (MMRL)

❖ Prompt Learning 의 발전과정

Text Prompt | **Multimodal Prompt** | **Multimodal Representation**

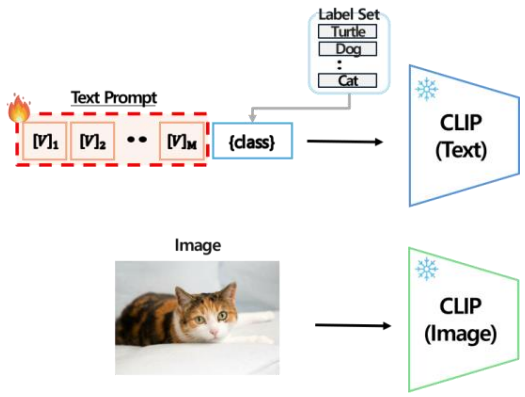
CoOp [IJCV 2022] → **CoCoOp** [CVPR 2022] → **MaPLe** [CVPR 2023] → **MMRL** [CVPR 2025]

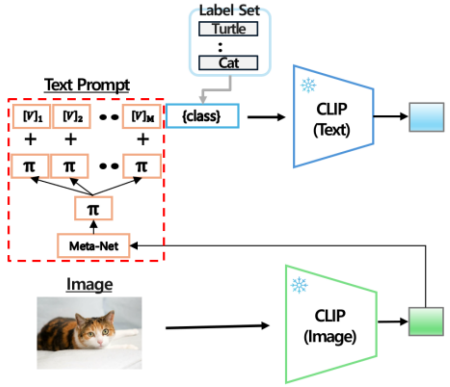
연속적인 Context vector 제안

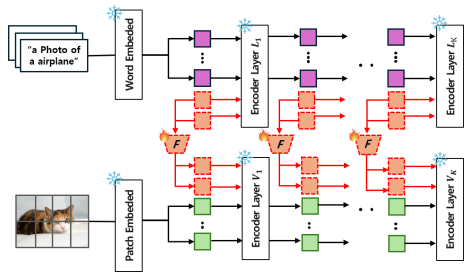
조건화된 Prompt Learning 제안

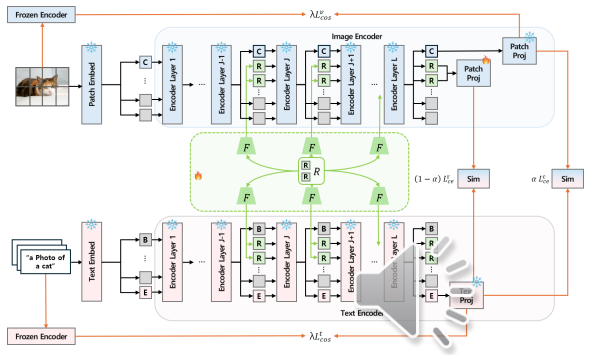
{Text + Visual}
Prompt Learning 제안

Multi-Modal Representation Learning
+ Decoupling Strategy









4. MMRL

Multi-Modal Representation Learning (MMRL)

❖ Prompt Learning 의 발전과정

Text Prompt | **Multimodal Prompt** | **Multimodal Representation**

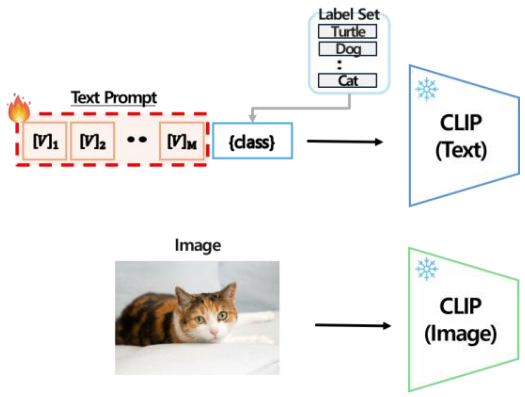
CoOp [IJCV 2022] → **CoCoOp** [CVPR 2022] → **MaPLe** [CVPR 2023] → **MMRL** [CVPR 2025]

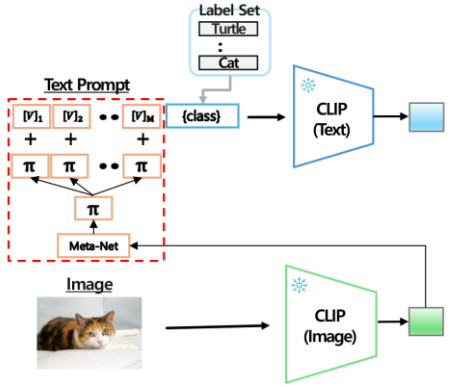
연속적인 Context vector 제안

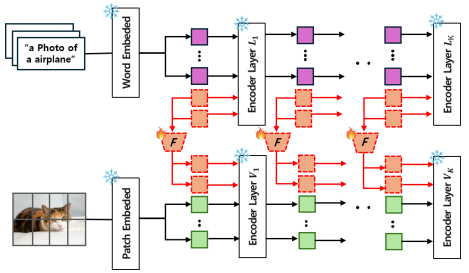
조건화된 Prompt Learning 제안

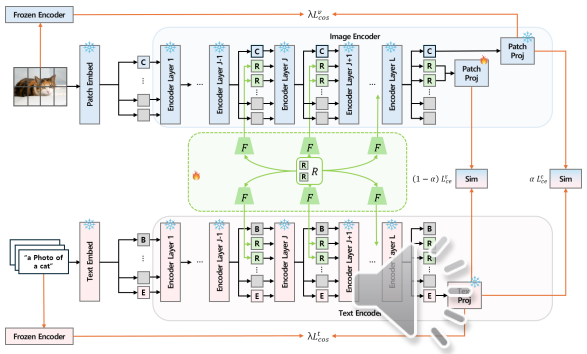
{Text + Visual}
Prompt Learning 제안

Multi-Modal Representation Learning
+ Decoupling Strategy









Thank You