2025 겨울 세미나

Prompt Learning with Knowledge Distillation

Pusan National University
Computer Vision & Signal Processing Lab, Dept. of Electronic Engineering

Presented By
백창우

CVSP Lab
Computer Vision & Signal Processing

Outline

- 주제 및 논문 선정 이유
- Background
  - Prompt Learning
  - Knowledge Distillation
- PromptKD: Unsupervised Prompt Distillation for Vision-Language Models (CVPR 2024) [1]
- Improving Zero-shot Generalization of Learned Prompts via Unsupervised Knowledge Distillations (ECCV 2024) [2]

주제 및 논문 선정 이유

- 주제: Prompt Learning with Knowledge Distillation
  - 2월부터 직접 연구 중
  - 쉽고 직관적인 개념, 실용성이 높아 선정
  - Prompt Learning은 VLM 기반 어플리케이션에서 활용도가 높아, 세미나를 듣는 청중에게 유용한 주제라 선정
  - 최근 Prompt Learning에 Knowledge Distillation을 접목시킨 연구가 등장하여 이를 소개하고자 함

- 논문 선정 이유
  - Prompt Learning with Knowledge Distillation을 짧은 시간 안에 효과적으로 설명할 수 있는 논문 선정
  - 간단하고 직관적으로 이해하기 쉬운 논문

Background

- Prompt Learning
  - Pretrained Vision-Language Model(VLM)에서 Prompt를 최적화하여 특정 Downstream Task에 적응하는 학습 기법
    - 주로 Zero Shot Classification 및 Retrieval Downstream Task에 적용
    - 모델을 파라미터를 업데이트 하지 않고도 적은 자원으로 성능 향상
    - 적은 데이터로도 학습 가능(Few- Shot Learning)
  - Prompt: 모델에 입력되는 어떤 문장 또는 토큰
    - Handcraft prompt: 사람이 직접 설계한 고정된 형태의 프롬프트
      - Ex) “A Photo of a Cat”
    - Soft Prompt: 모델이 학습을 통해 최적화한 연속형 벡터 기반의 프롬프트
      - CoOp [1]

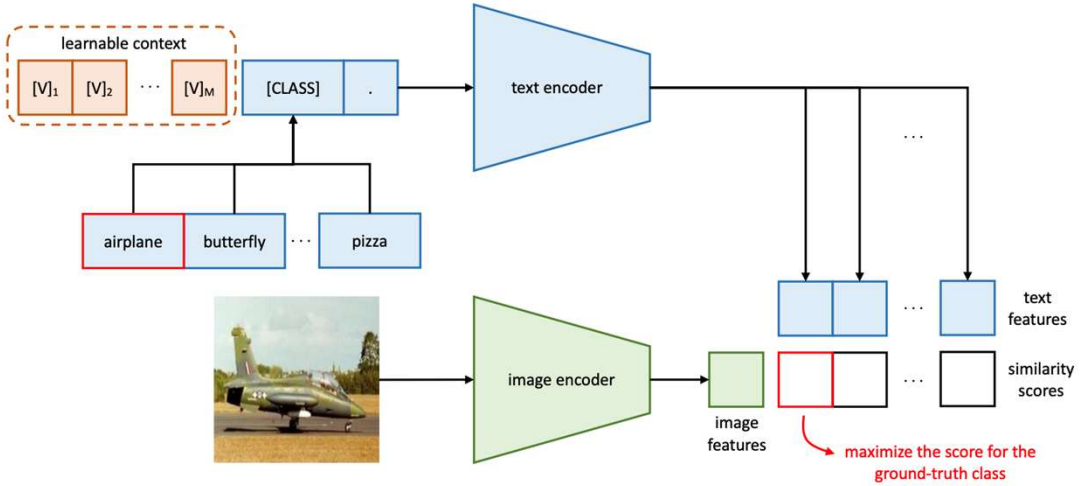


Figure 2: Overview of context optimization (CoOp).

Background

- Prompt Learning
  - Text Prompt Learning
    - Text 기반 Prompt를 학습하는 방식
  - Visual Prompt Learning
    - 이미지 인코더(ViT) 입력에 학습 가능한 Visual Prompt를 삽입하여 모델 성능을 조정하는 방식
  - Text-Visual Prompt Learning
    - Text Prompt와 Visual Prompt를 동시에 함께 학습하여 멀티 모달 정보 간의 상호작용을 최적화하는 방식

Background

- Prompt Learning
  - Text Prompt Learning
    - Text 기반 Prompt를 학습하는 방식
      - Inference시에 고정된 Softprompt 방식
        - ✓Unseen Class에서 일반화 성능이 낮음
        - ✓CoOp [3]
      - Inference시에 이미지의 입력에 따라 바뀌는 동적인 Softprompt 방식
        - ✓CoCoOp [4]

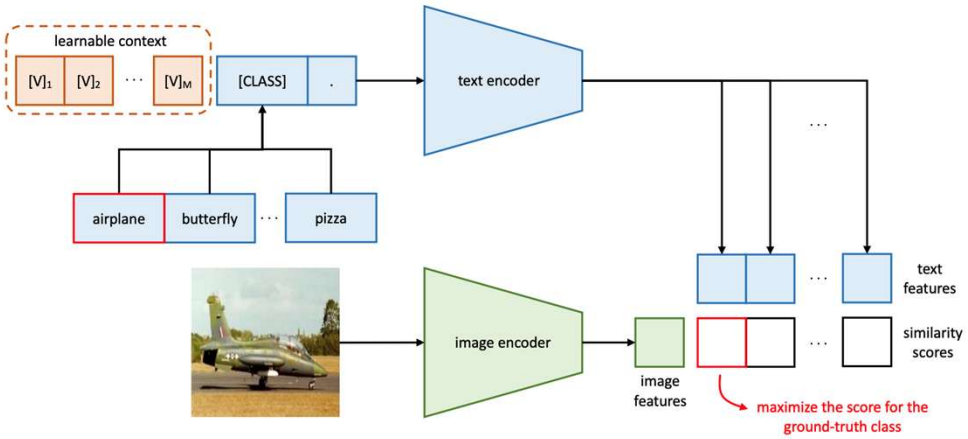


Figure 2: Overview of context optimization (CoOp).

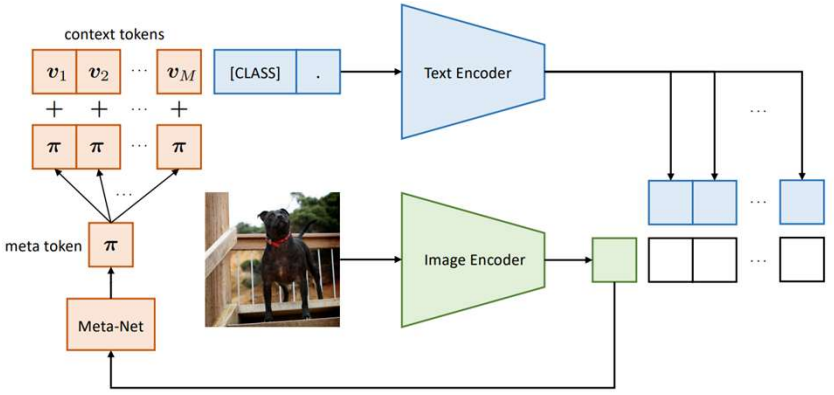
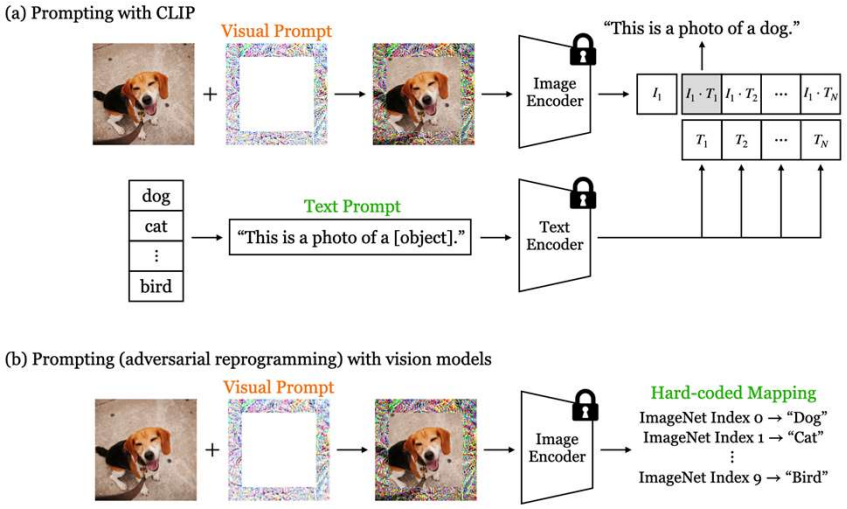
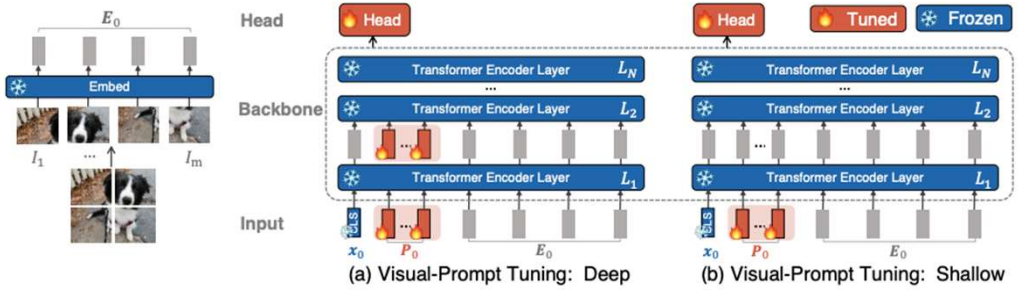


Figure 2. Our approach, Conditional Context Optimization (CoCoOp), consists of two learnable components: a set of context vectors and a lightweight neural network (Meta-Net) that generates for each image an input-conditional token.

Background

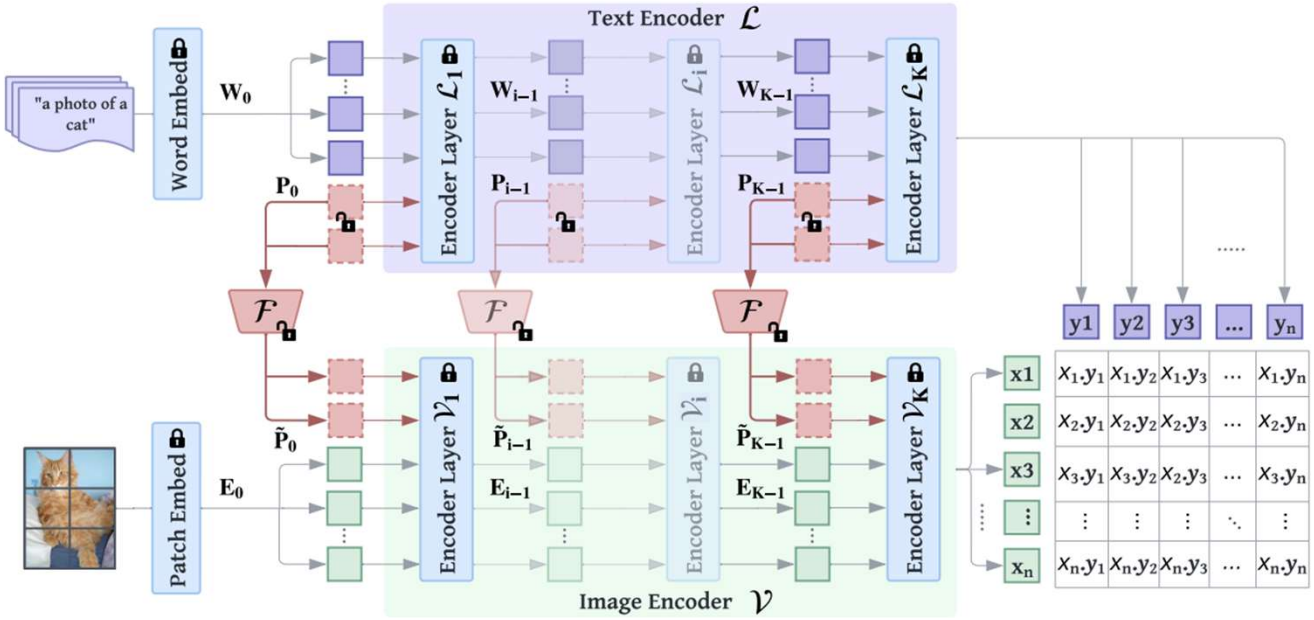
- Prompt Learning
  - Visual Prompt Learning
    - 이미지 인코더(ViT) 입력에 학습 가능한 Visual Prompt를 삽입하여 모델 성능을 조정하는 방식
      - Fixel Level에서 액자 형태의 Prompt를 추가하는 방식
        - ✓Exploring Visual Prompts for Adapting Large-Scale Models [5]
      - Text Prompt 처럼 ViT 입력에 Leanable한 Token을 추가하는 방식
        - ✓Visual Prompt Tunning [6]





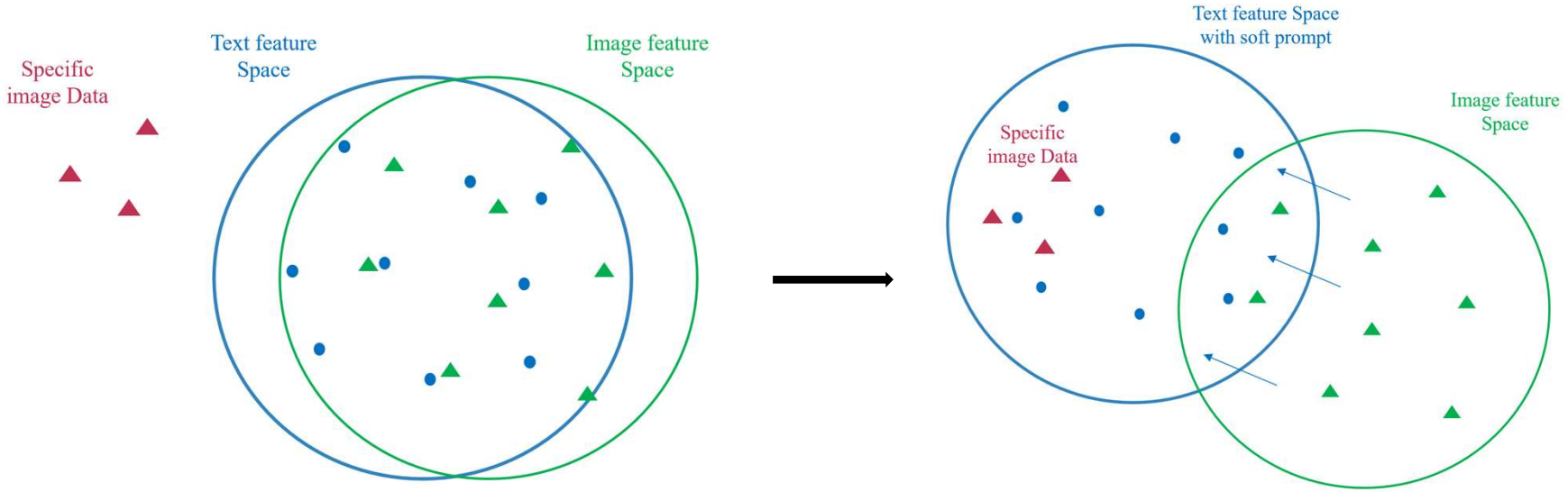
Background

- Prompt Learning
  - Text-Visual Prompt Learning
    - Text Prompt와 Visual Prompt를 동시에 함께 학습하여 멀티 모달 정보 간의 상호작용을 최적화하는 방식
      - Text와 Image Space를 유연하게 동적으로 조정하여 일반화에 유리
        - ✓MaPLe: Multi-modal Prompt Learning [7]



Surplus Material

- Prompt Learning Weakness
  - Soft Prompt가 특정 Downstream Task 데이터에 Overfitting될 가능성이 높음
    - 새로운 도메인이나 미지의 데이터에서 일반화 성능이 저하될 우려
    - 이를 해결하고자 Out-of-Distribution이나 Domain Generalization 등 Prompt Learning의 Robustness 연구가 최근 활발히 진행 중
      - 성능 지표 또한 DG, Cross-Dataset, Unseen Class 등에 대해 주로 평가됨

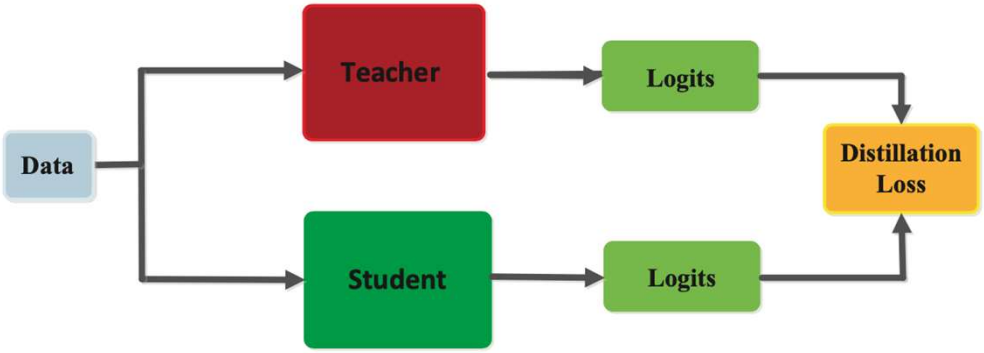


CLIP Feature Space

After Prompt Learning

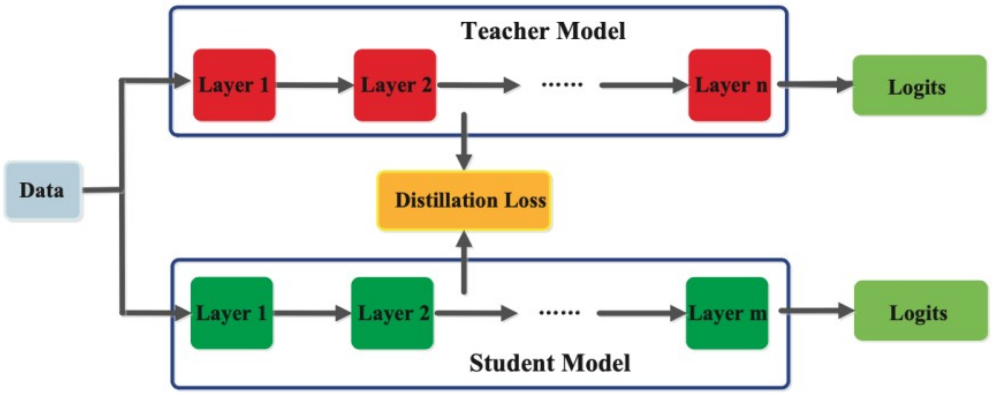
Background

- Knowledge Distillation
  - 미리 잘 학습된 큰 네트워크(Teacher Network)의 지식을 실제로 사용하고자 하는 작은 네트워크(Student Network)에 전달하는 경량화 기법
    - Logit Distillation
      - Teacher 모델의 출력 확률 분포(soft labels)를 Student 모델이 학습하도록 하는 방식
    - Feature Distillation
      - Teacher 모델의 중간 및 최종 feature representation을 Student 모델이 학습하도록 하는 방식



$$\mathcal{L}_{\mathrm{KD}} = \alpha \times H\big(y, \sigma(z_s)\big) + \beta \times H\big(\sigma(z_t;\rho), \sigma(z_s;\rho)\big)$$

GT → Student Teacher → Student



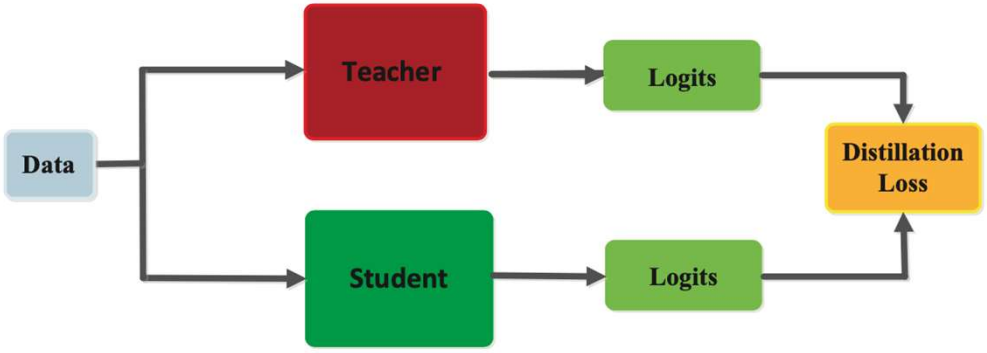
$$\mathcal{L}_{\mathrm{KD}} = \alpha \times H\big(y, \sigma(z_s)\big) + \beta \times \|f_T - f_S\|^2$$

GT → Student Teacher → Student

CVSP Lab
Computer Vision & Signal Processing

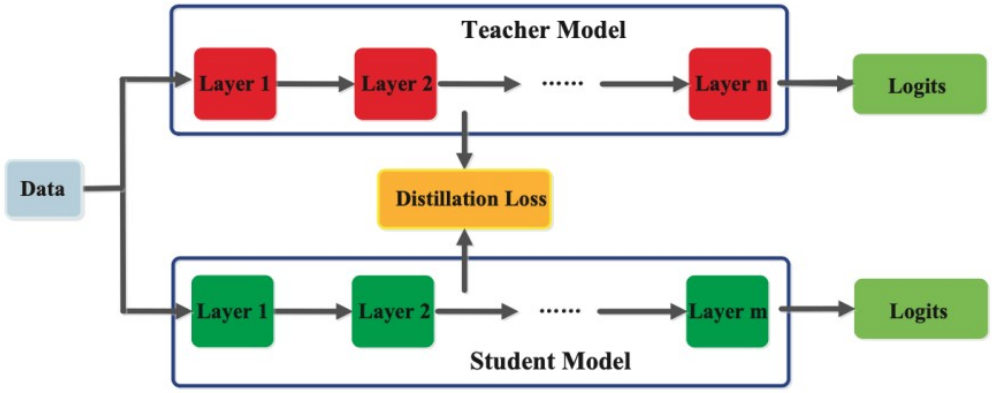
Background

- Unsupervised Knowledge Distillation
  - 라벨이 없는 데이터(Unlabeled Data)에서 Knowledge Distillation을 수행하는 방식
    - Teacher 모델이 제공하는 예측 값(Soft Target)만을 사용하여 Student 모델을 학습
      - Teacher 모델의 출력 분포(Soft Label) 자체를 학습 목표로 삼음
      - Student 모델이 Teacher가 가진 Feature Representation과 지식을 모방하도록 유도
      - Teacher 모델의 출력을 바탕으로 Pseudo Label을 생성하여 학습하는 방식도 존재



$$\mathcal{L}_{KD} = \alpha \times H(y, \sigma(z_s)) + \beta \times H(\sigma(z_t;\rho), \sigma(z_s;\rho))$$

GT → Student Teacher → Student



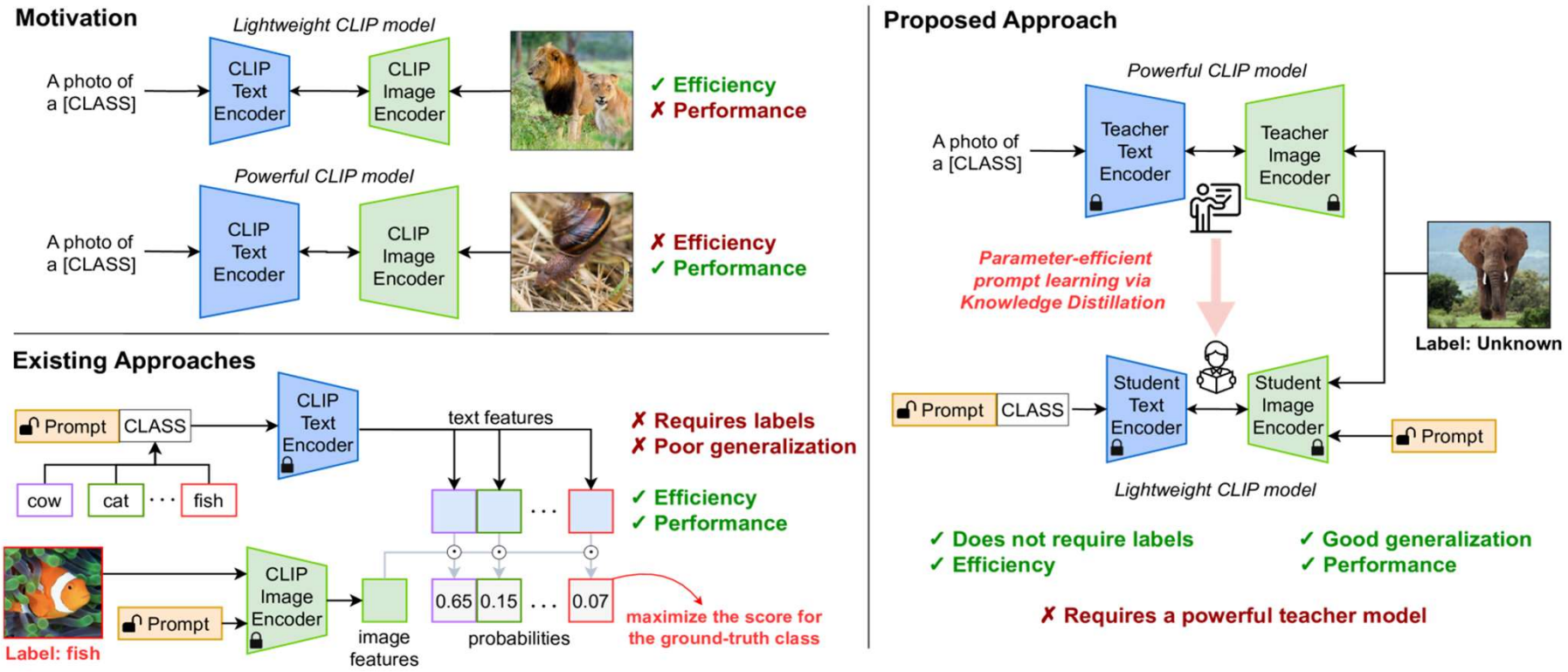
$$\mathcal{L}_{KD} = \alpha \times H(y, \sigma(z_s)) + \beta \times \|f_T - f_S\|^2$$

GT → Student Teacher → Student

CVSP Lab Computer Vision & Signal Processing

Background

- Task Overview
  - Prompt Learning with Knowledge Distillation
    - Large VLM에서 Small VLM으로 지식을 전달할 때 Prompt를 학습하여 Knowledge Distillation을 수행하는 방법
      - 모델 Parameter를 업데이트하지 않고도 Knowledge Distillation 가능
      - Teacher 모델이 강력한 일반화 성능을 제공하여 Prompt Learning의 일반화 문제 해결 가능
      - Teacher Model을 두어 Unsupervised 방식으로도 학습 가능
        - ✓Pseudo Label 효과



PromptKD: Unsupervised Prompt Distillation for Vision-Language Models

CVPR 2024

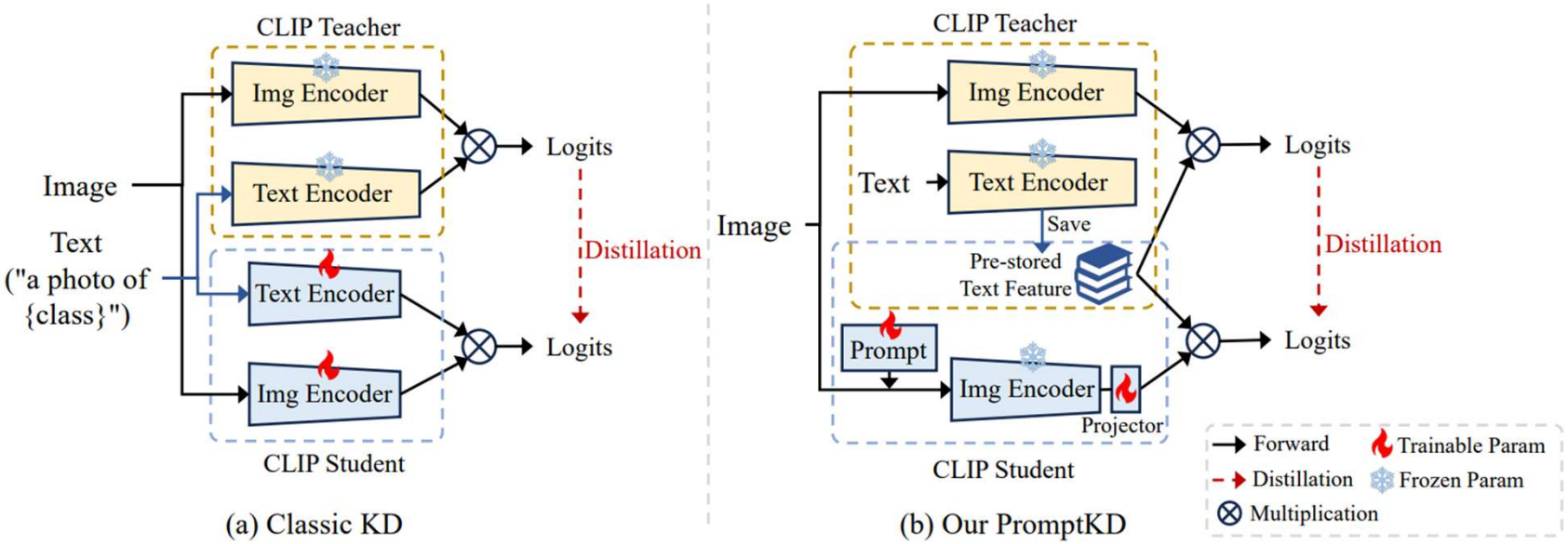
Introduction

- Motivation
  - 기존 CLIP 기반 Prompt Learning 방식은 Label이 필요

    => KD 패러다임을 통해 Label이 없는 상황에서 학습
  - 최초로 Unsupervised Prompt KD 방식 제안

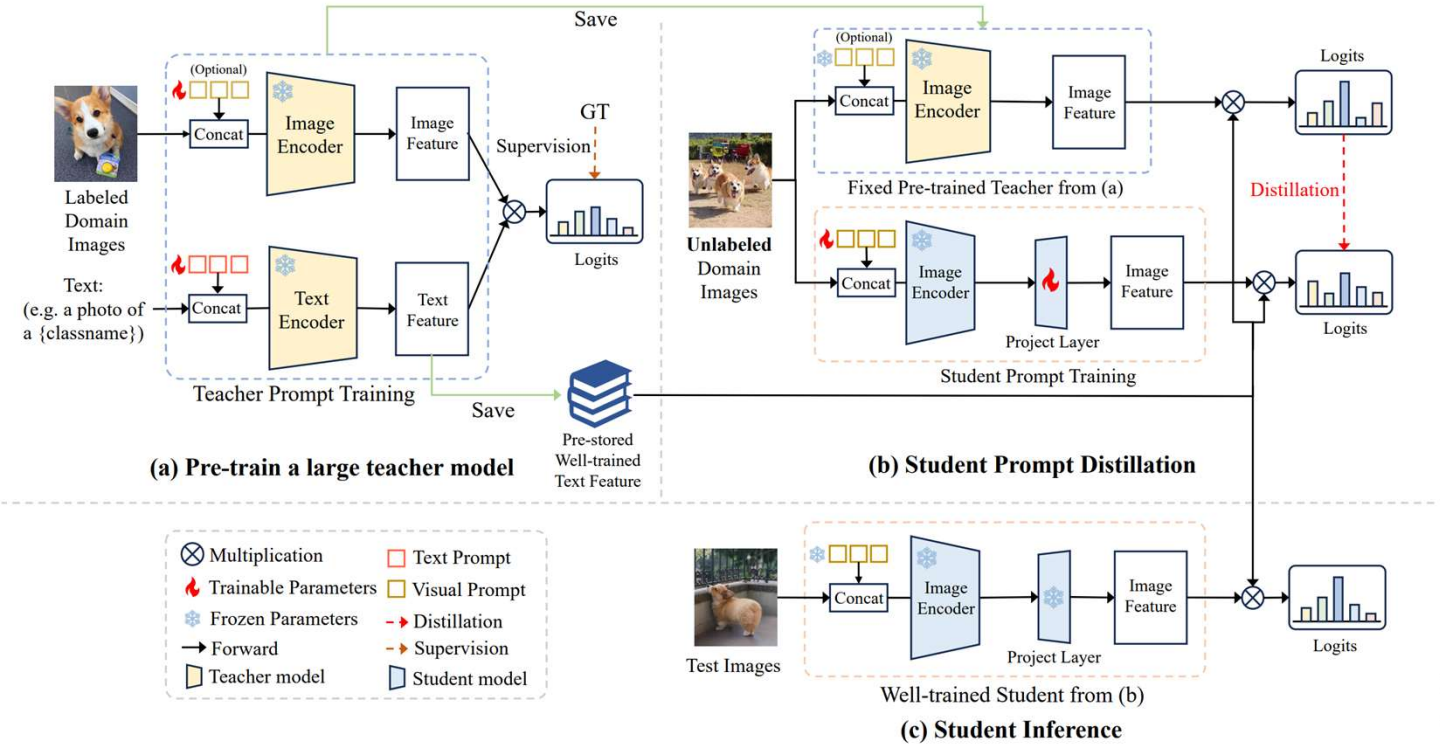
    - 기존 KD 방식과 달리 모델의 Parameter 업데이트 하지 않고 Prompt와 Projector만 학습하는 효율적인 방식

    - Teacher Text Encoder의 Feature를 저장하는 방식으로 Inference 연산 비용 감소



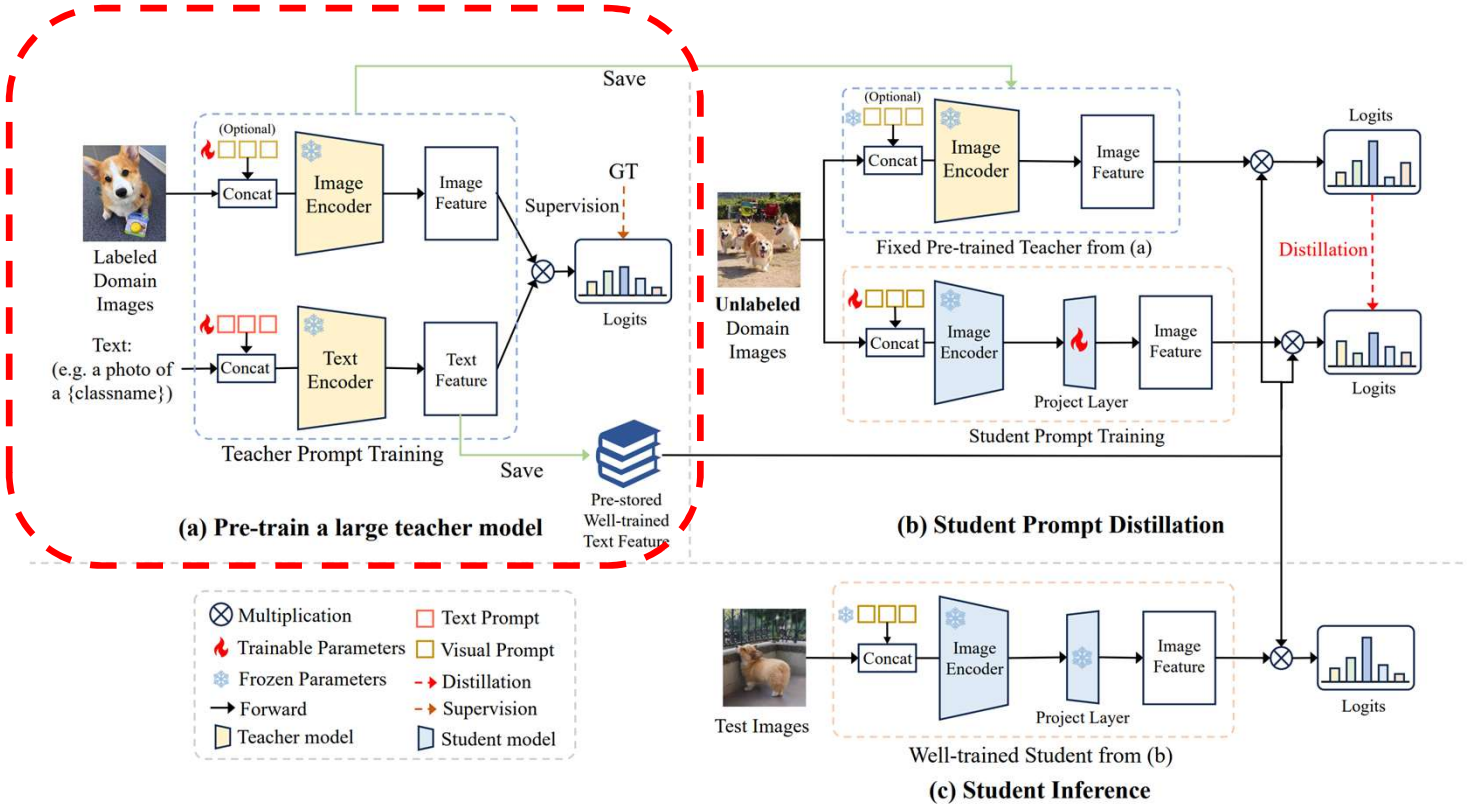
Method

- Method overview
  - Stage1: Label이 있는 도메인 데이터로 Supervised Learning
    - Teacher Text Prompt를 학습 후 Text feature를 저장
  - Stage2: Label이 없는 도메인 데이터로 Unsupervised Learning
    - **Teacher CLIP(prompt trained)의 Logits과 저장한 Teacher Text feature와 Student Image feature 간 Logits**을 KL Divergence Loss로 Distillation
      - Student image encoder의 visual prompt와 Project Layer를 학습
  - Inference: Teacher Text Feature와 Student Image feature간 Logits으로 Classify



Method

- Method
  - Stage1
    - 라벨이 있는 도메인 데이터를 사용하여 Teacher CLIP 모델을 Prompt Learning 방식으로 사전 학습
      - Prompt: [Trainable Soft Prompts], 'a photo of a {class name}'
      - Cross Entropy Loss로 학습
    - Prompt 학습이 완료되면 Text Encoder 출력을 저장
      - Student text encoder를 사용하지 않아 학습 비용 절감

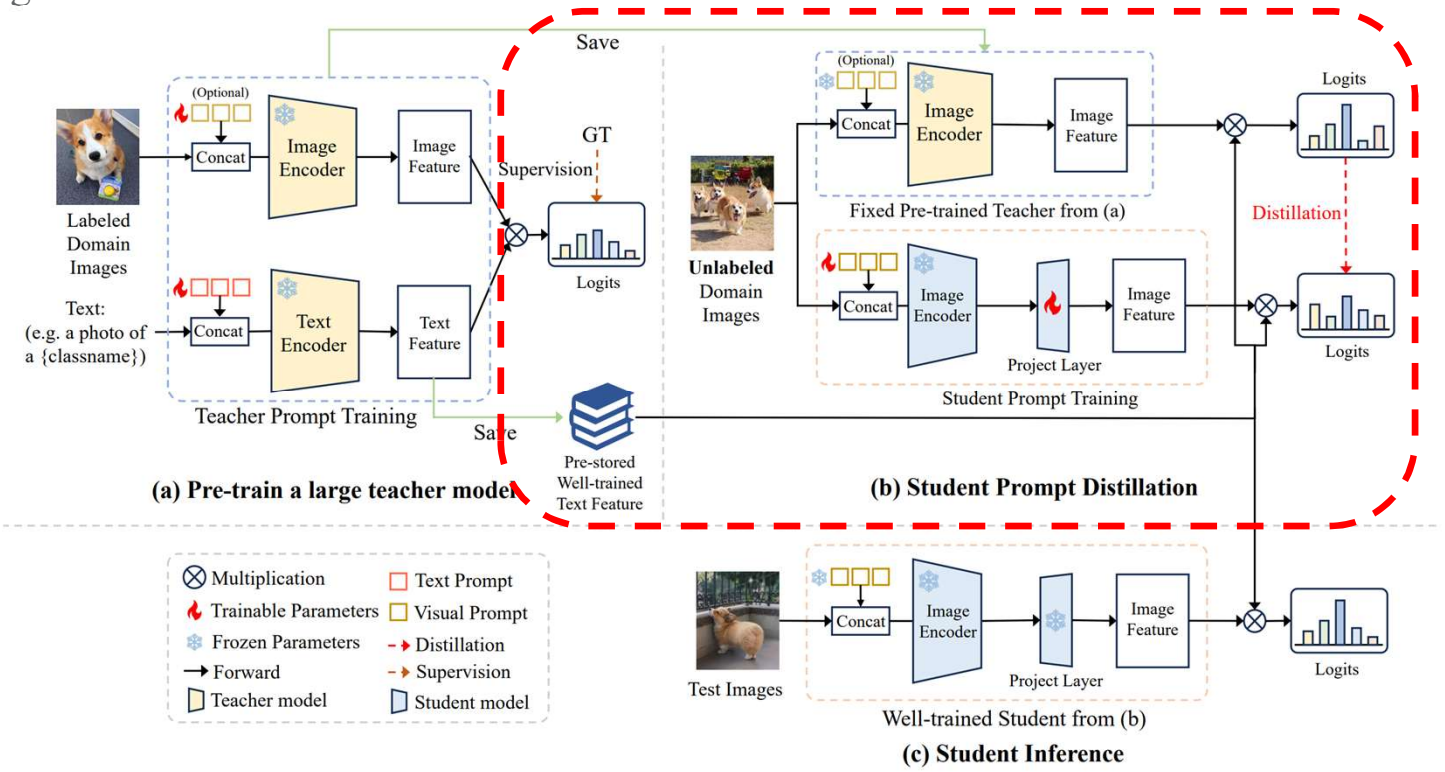


CVSP Lab
Computer Vision & Signal Processing

Method

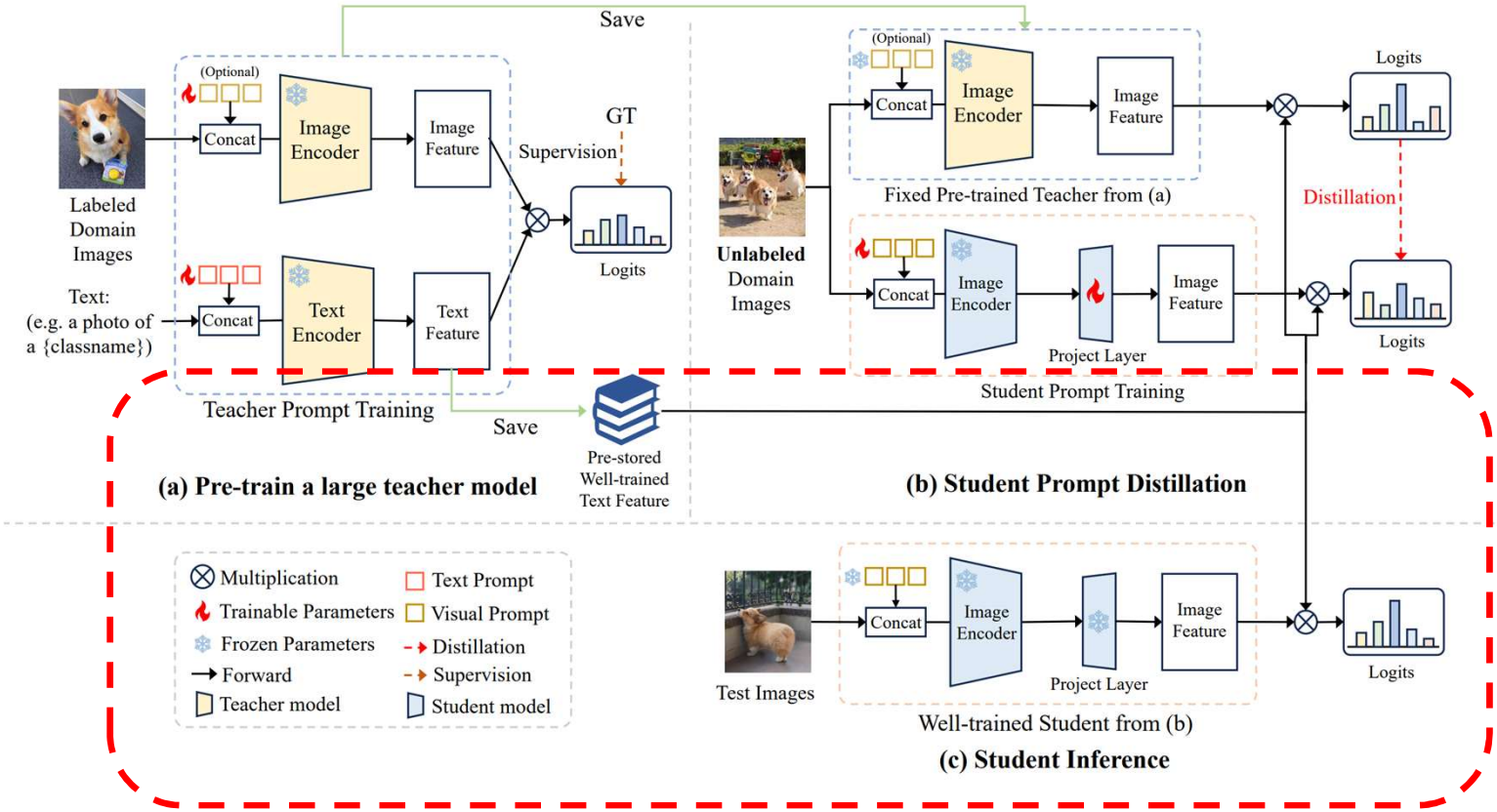
- Method
  - Stage2
    - 라벨이 없는 도메인 데이터를 사용하여 Unsupervised Logits Distillations
      - Teacher Network Logits: 저장된 Text prompt feature와 Teacher image Encoder feature간 Logits
      - Student Network Logits: 저장된 Text prompt feature와 Student image Encoder간 Logits
      - KL Divergence Loss로 두 Logits을 Distillations
        - ✓q^t: Teacher Image feature
        - ✓q^s: Student Image feature

$$L_{kd}(q^t, q^s, \tau) = \tau^2 KL(\sigma(q^t/\tau), \sigma(q^s/\tau))$$



Method

- Method
  - Inference
    - 저장된 Text prompt feature와 Student image Encode간 Logits으로 Classify



Experiments

- Downstream Task
  - Zero-Shot Classification(few-shot learning으로 진행)
    - Base-to-novel Generalization
      - 기존(Base) 클래스에서 학습한 모델이 새로운(Novel) 클래스로 일반화되는 능력을 평가
        - ✓Train set을 Base Class와 Novel Class로 나누어 Base Class로만 학습하고 Novel(Unseen) Class에 대해 성능을 측정
    - Cross-dataset Evaluation
      - 특정 Source Dataset에서 학습한 모델을 다른 Target Dataset에서 성능 평가
        - ✓학습 모델의 일반화 능력 평가
        - ✓본 논문에선 Unsupervised learning(Stage2)으로 Target Dataset을 학습

Experiments

- Zero Shot Classification(Base-to-novel Generalization)
  - Teacher 및 Student Network
    - Teacher Model: VIT-L/14 CLIP(PromptSRC [8])
    - Student Model: VIT-B/16 image encoder of the CLIP
  - 16 Shot

Table 2. Configuration of paired visual and text encoders.

Visual encoder			Text encoder: Transformer [42]			
Model	Type	Params	Layer	Width	Head	Params
ViT-L/14 [8]	ViT	304.0M	12	768	12	85.1M
ViT-B/16 [8]		86.2M	12	512	8	37.8M
ViT-T/16 [8]		5.6M	12	384	6	21.3M
MobileViT-S [32]		5.3M				
Swin-T [30]		27.9M				
ResNet-101 [12],	CNN	56.3M	12	512	8	37.8M
ResNet-50 [12],		38.3M				
ResNet-18 [12],		11.4M	12	384	6	21.3M
MobileNetV3 [16]		2.0M				
EfficientNet-B0 [41]		4.7M				

ViT-B/16	Base	Novel	HM
CLIP	69.34	74.22	71.70
CoOp	82.69	63.22	71.66
CoCoOp	80.47	71.69	75.83
MaPLe	82.28	75.14	78.55
PromptSRC	84.26	76.10	79.97
PromptKD	86.96	80.73	83.73
Δ	+2.70	+4.63	+3.76

(a) Average over 11 datasets.

ViT-B/16	Base	Novel	HM
CLIP	72.43	68.14	70.22
CoOp	76.47	67.88	71.92
CoCoOp	75.98	70.43	73.10
MaPLe	76.66	70.54	73.47
PromptSRC	77.60	70.73	74.01
PromptKD	80.83	74.66	77.62
Δ	+3.23	+3.93	+3.61

(b) ImageNet

ViT-B/16	Base	Novel	HM
CLIP	96.84	94.00	95.40
CoOp	98.00	89.81	93.73
CoCoOp	97.96	93.81	95.84
MaPLe	97.74	94.36	96.02
PromptSRC	98.10	94.03	96.02
PromptKD	98.91	96.65	97.77
Δ	+0.81	+2.62	+1.75

(c) Caltech101

ViT-B/16	Base	Novel	HM
CLIP	91.17	97.26	94.12
CoOp	93.67	95.29	94.47
CoCoOp	95.20	97.69	96.43
MaPLe	95.43	97.76	96.58
PromptSRC	95.33	97.30	96.30
PromptKD	96.30	98.01	97.15
Δ	+0.97	+0.71	+0.85

(d) OxfordPets

ViT-B/16	Base	Novel	HM
CLIP	63.37	74.89	68.65
CoOp	78.12	60.40	68.13
CoCoOp	70.49	73.59	72.01
MaPLe	72.94	74.00	73.47
PromptSRC	78.27	74.97	76.58
PromptKD	82.80	83.37	83.13
Δ	+4.53	+8.40	+6.55

(e) StanfordCars

ViT-B/16	Base	Novel	HM
CLIP	72.08	77.80	74.83
CoOp	97.60	59.67	74.06
CoCoOp	94.87	71.75	81.71
MaPLe	95.92	72.46	82.56
PromptSRC	98.07	76.50	85.95
PromptKD	99.42	82.62	90.24
Δ	+1.35	+6.12	+4.29

(f) Flowers102

Table 1. Comparison with existing state-of-the-art methods on base-to-novel generalization. Our proposed PromptKD demonstrates strong generalization ability and achieves significant improvements on 11 recognition datasets given the **ViT-B/16 image encoder** of the CLIP model. In our approach, the default teacher model is the ViT-L/14 CLIP model. The symbol Δ denotes the performance improvement compared to the previous SOTA method PromptSRC. Our PromptKD outperforms previous methods on all datasets.

CVSP Lab
Computer Vision & Signal Processing

Experiments

- Zero Shot Classification(Cross Dataset Evaluation)
  - Teacher 및 Student Network
    - Teacher Model: VIT-L/14 CLIP(PromptSRC [8])
    - Student Model: VIT-B/16 image encoder of the CLIP
  - Source Dataset: ImageNet(16-Shot)
    - Target Datasets: Caltech101, Oxford Pets, Standford Cars, Flowers102 등 총 10개의 Datasets
      - Stage2 Unsupervised 방식으로 학습

		Target Dataset										
ZSL	ViT-B/16	Caltech 101	Oxford Pets	Standford Cars	Flowers 102	Food101	FGVC Aircraft	SUN397	DTD	Euro SAT	UCF101	Avg.
In-ductive	CoOp	93.70	89.14	64.51	68.71	85.30	18.47	64.15	41.92	46.39	66.55	63.88
	CoCoOp	**94.43**	90.14	65.32	71.88	86.06	22.94	67.36	45.73	45.37	68.21	65.74
	MaPLe	93.53	90.49	65.57	72.23	86.20	24.74	67.01	46.49	48.06	68.69	66.30
	PromptSRC	93.60	90.25	65.70	70.25	86.15	23.90	67.10	46.87	45.50	68.75	65.81
Trans-ductive	PromptKD	93.61	**91.59**	**73.93**	**75.33**	**88.84**	**26.24**	**68.57**	**55.08**	**63.74**	**76.39**	**71.33**
	Δ	+0.01	+1.34	+8.23	+5.08	+2.69	+2.34	+1.47	+8.21	+18.24	+7.64	+5.52

Table 2. Comparison of PromptKD with existing advanced approaches on cross-dataset benchmark evaluation. Based on our pipeline, we perform unsupervised prompt distillation using the unlabeled domain data respectively (i.e., the transductive setting). The source model is trained on ImageNet [7]. "ZSL" denotes the setting type for Zero-Shot Learning. PromptKD achieves better results on 9 of 10 datasets.

CVSP Lab
Computer Vision & Signal Processing

Conclusion

- Pros
  - 최초의 Unsupervised Knowledge Prompt Distillation
  - Teacher text feature를 저장하여 Inference 연산 비용 감소
  - 높은 성능 향상

- Cons
  - 학습 시 두 종류의 Data 필요
    - Stage1에서 Data Label이 필요
      - 완전한 의미의 Unsupervised Learning은 아님
  - Prompt만 학습하는 것이 아니라 추가적인 Projector 필요
    - 추가적인 연산 비용 발생

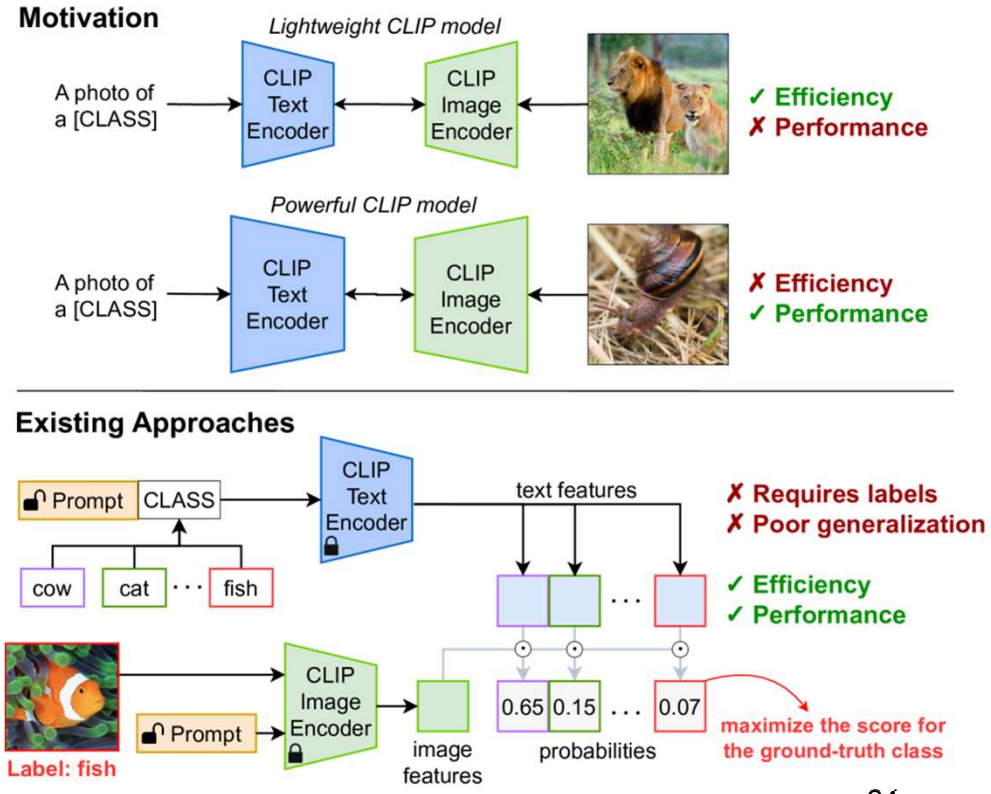
Improving Zero-shot Generalization of Learned Prompts via Unsupervised Knowledge Distillations

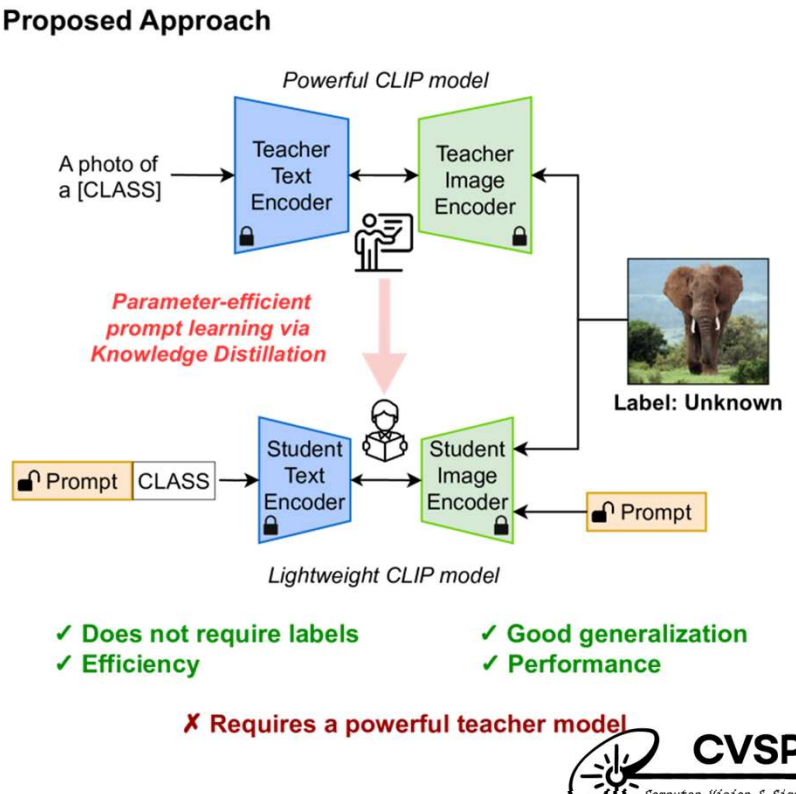
ECCV 2024

CVSP Lab
Computer Vision & Signal Processing

Introduction

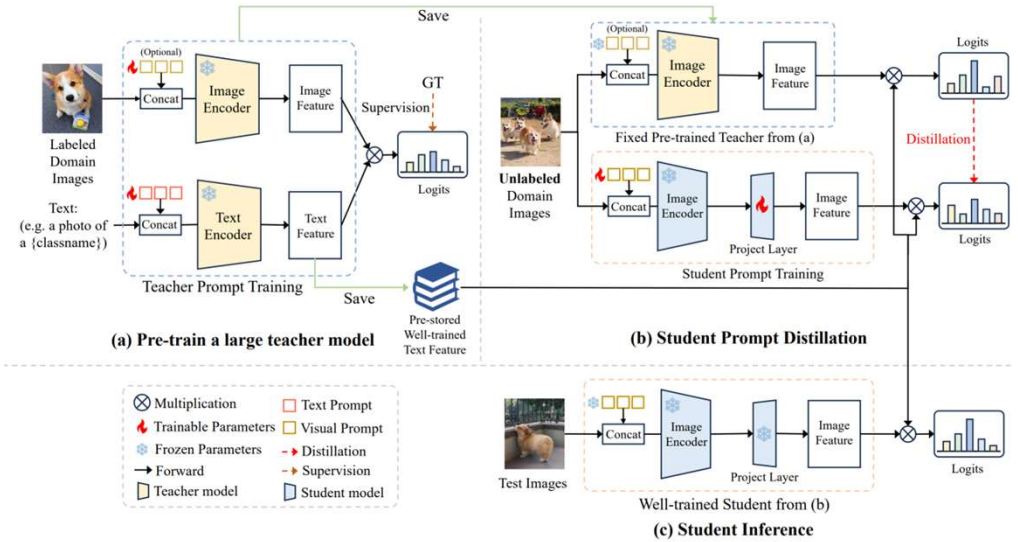
- Motivation
  - 기존 Prompt Learning 방식들이 Downstream Task에 Overfitting되어 일반화 성능이 떨어지는 문제를 일반화 성능이 강력한 Large CLIP 모델을 활용하여 해결
  - Label Free 방식으로 Unsupervised 방식으로 학습
    - 데이터셋의 Class 리스트조차 모르는 Class-Agnostic 상황에서도 학습 가능
  - CoOp, VPT, Maple 등 여러 Prompt Learninig 기법에 적용 가능한 범용적인 Plugin KD

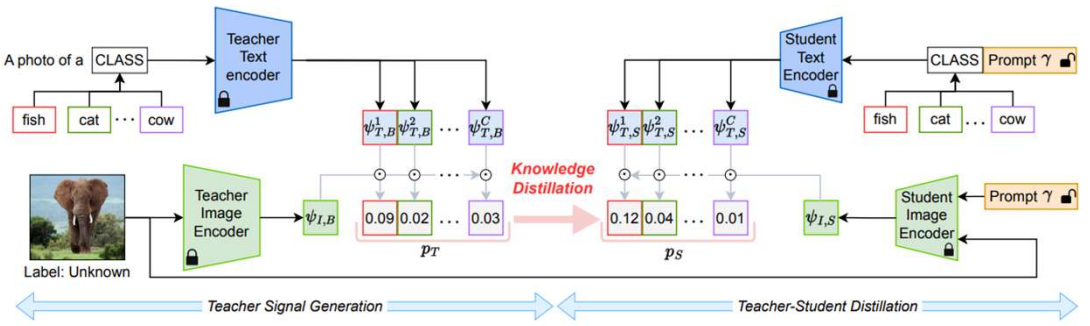




Introduction

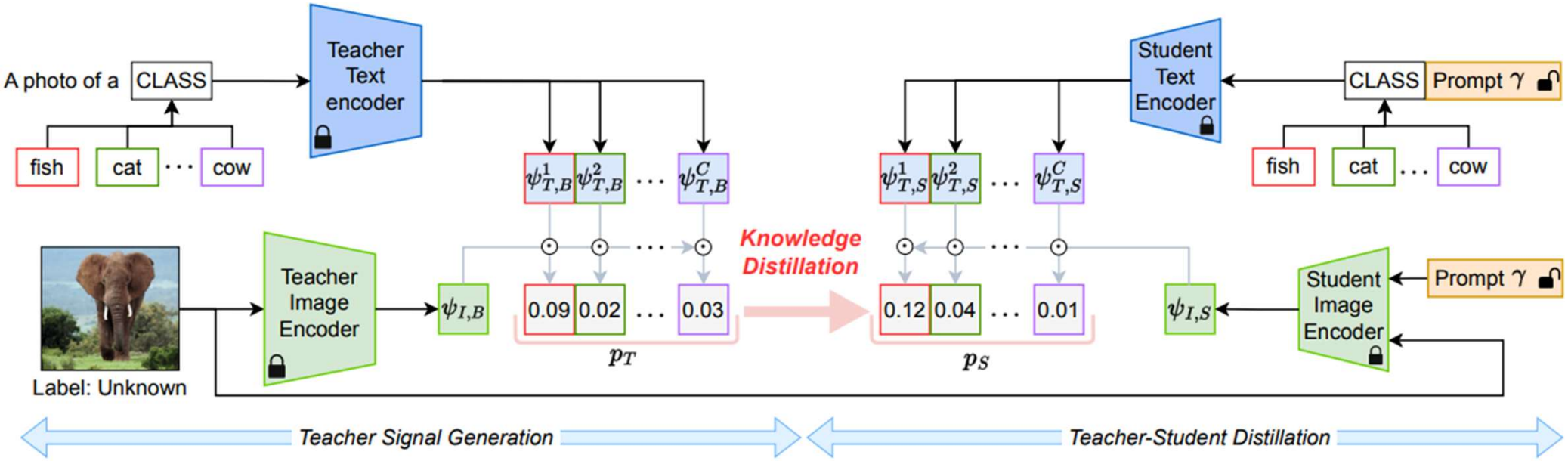
- Motivation
  - PromptKD Vs Ours(KDPL)
    - 1. PromptKD는 Teacher Model을 Supervised 방식으로 사전 학습해야 함
    - 2. PromptKD는 Project Layer를 필요로 하므로 파라미터가 추가됨
    - 3. PromptKD는 두가지 데이터를 학습에 사용하지만 KDPL(ours)는 표준 16-Shot 설정을 따름





Method

- Method Overview
  - Unsupervised Logits Distillation 방식으로 학습
    - Teacher CLIP의 Logits과 Student CLIP의 Logits간 KL Divergence Loss로 Distillation
  - Class agnostic learning
    - Dataset의 Class 리스트조차 안주어진 상황
      - Class 리스트를 고르기 위해 20k의 Class 후보 중 Teach model의 Logits 기준으로 상위 Top K Class로 설정
        - ✓Class 수에 따라 메모리 요구량은 선형적으로 증가하므로 20k Class 전부 학습하는 것은 불가능(300GB의 메모리 요구)

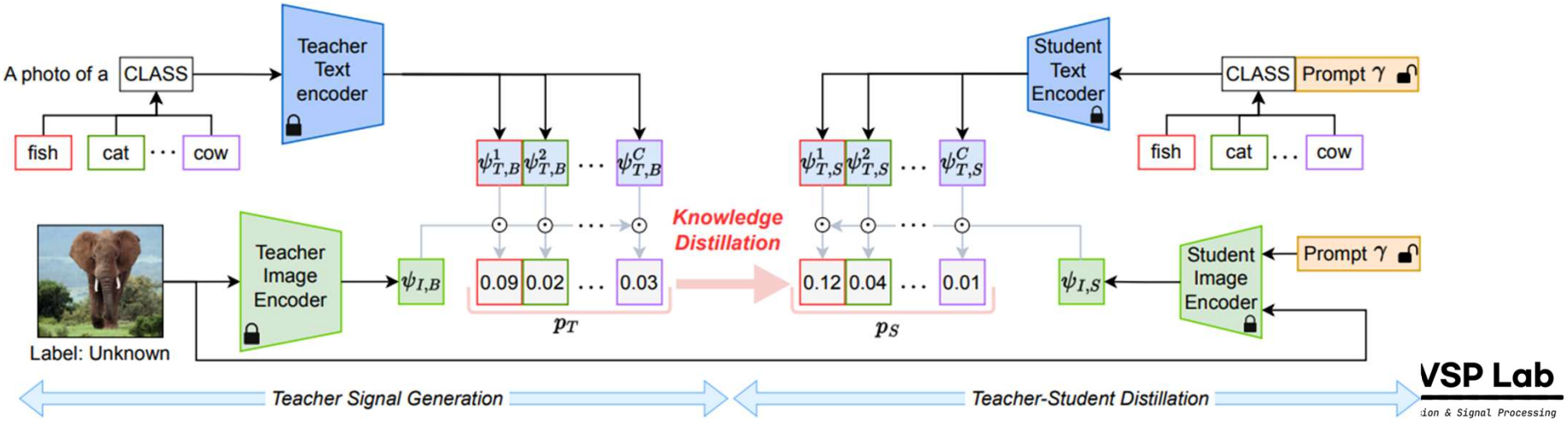


Method

- Unsupervised Logits Distillation
  - Training
    - Teacher Model의 Logits과 Student Model의 Logits간 symmetric-KL Divergence Loss를 통해 학습
      - CoOp, VPT, MaPle 등 여러 모델 구조에 맞게 Text Prompt와 Visual Prompt만을 학습
      - Teacher Model은 Hand-craft Prompt를 사용
    - Train Loss
      - I: 이미지
      - $C = \{CLASS_i\}_{i=1}^{C}$: Class 집합
      - $p_S(I, C)$: Student Model의 Logits
      - $p_T(I, C)$: Teacher Model의 Logits

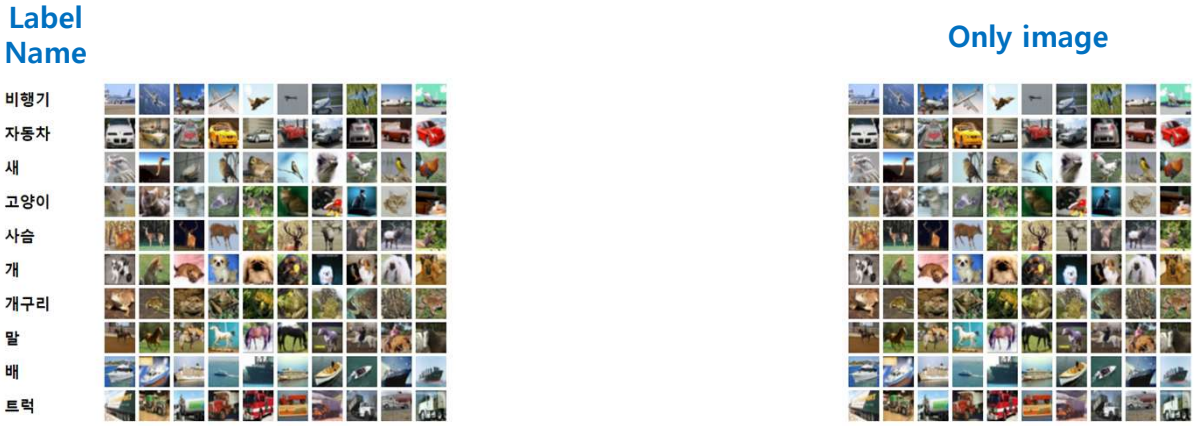
$$\mathcal{L}(\mathcal{I}, \mathcal{C}) = [D_{\mathrm{KL}}(p_T(I, \mathcal{C}) \,||\, p_S(I, \mathcal{C})) + D_{\mathrm{KL}}(p_S(I, \mathcal{C}) \,||\, p_T(I, \mathcal{C}))]$$

$$D_{\mathrm{KL}}(p \,||\, q) = \sum_{i \in \mathcal{C}} p_i \log\left(\frac{p_i}{q_i}\right)$$



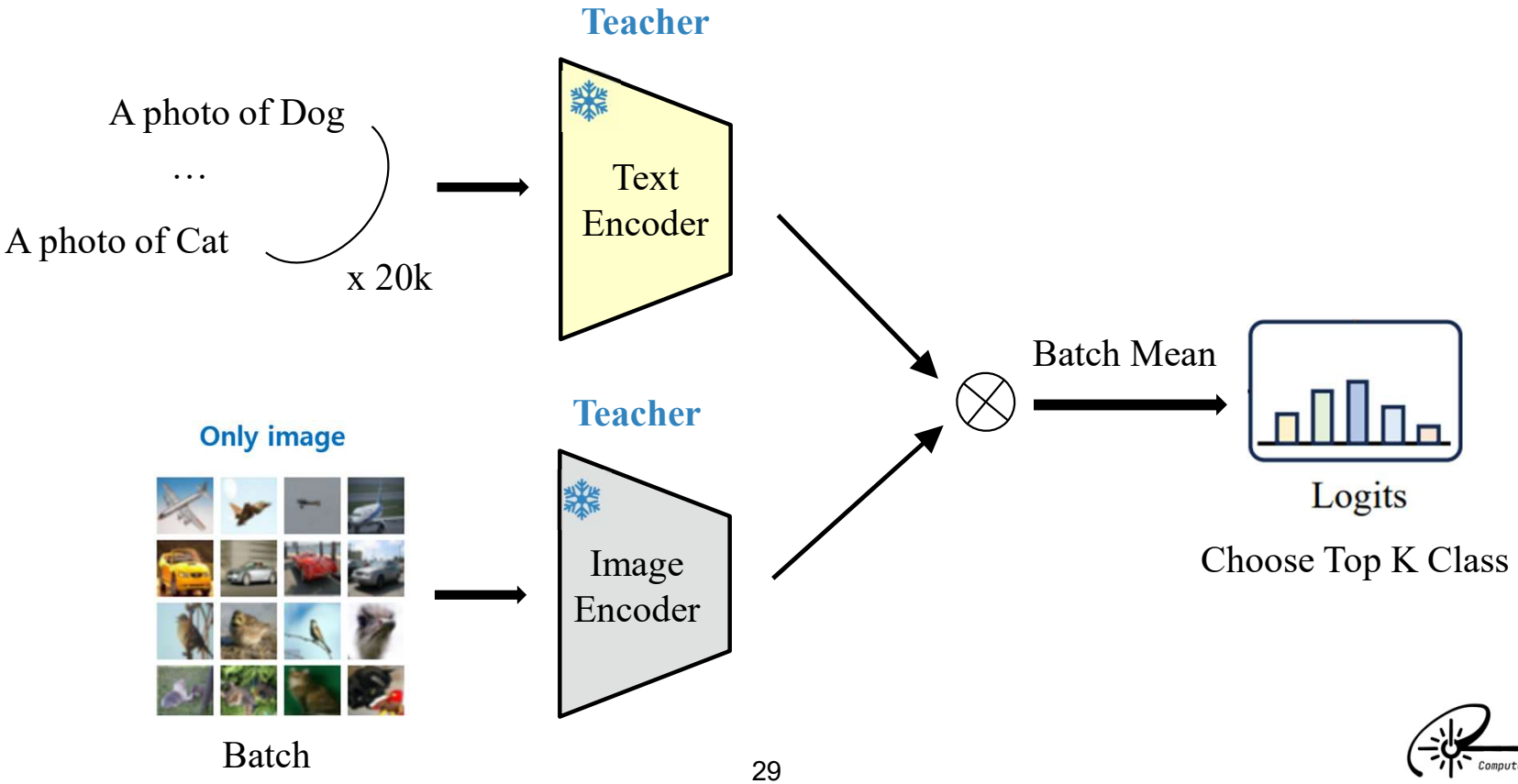
Method

- Class agnostic learning
  - Assumed Setting
    - 기존 Unsupervised 학습 환경
      - 개별 데이터 샘플에 대한 Label이 제공되지 않는 상황
        - ✓데이터셋 전체의 Class 리스트는 주어지지만, 각 샘플의 Class 정보는 없음
    - 본 논문에서 가정하는 Class Agnostic 학습 환경
      - 개별 샘플에 대한 Label 뿐만 아니라 Class 리스트도 주어지지 않은 상황
        - ✓오직 이미지 데이터만 주어진 상태에서 학습 진행



Method

- Class agnostic learning
  - Choose Class List
    - Class List를 고르기 위해 20k의 Class 후보 중 Teach model의 Logits 기준으로 상위 Top K Class로 설정
      - 20k의 Class 후보는 Open Images V7 데이터셋의 Clas List 사용
      - Batch 단위로 Class List 선정
        - ✓Image 별 Logits을 Batch 단위에서 평균을 낸 후 K개 Class 선택



Experiments

- Downstream Task
  - Zero-Shot Classification(few-shot learning으로 진행)
    - Domain Generation
      - Source Domain에서 학습 후 Target Domain에서 성능 평가
    - Cross Dataset Transfer(Evaluation)
      - 기존(Base) 클래스에서 학습한 모델이 새로운(Novel) 클래스로 일반화되는 능력을 평가
    - Generalization to Unseen Classes(Base-to-novel Generalization)
      - 특정 Source Dataset에서 학습한 모델을 다른 Target Dataset에서 성능 평가

Experiment

- Zero Shot Classification(Domain Generalization)
  - Base Model
    - 성능 비교 모델인 CoOp, VPT 등 Base모델들은 Supervised 방식으로 학습
  - Source Data
    - ImageNet1k
  - Target Data
    - ImageNet-V2
    - ImageNet-S
    - ImageNet-A
    - ImageNet-R

Table 1: Domain Generalization. Comparison between the baselines and the proposed unsupervised KDPL variants (highlighted in cyan). In this setting, the prompt is trained on ImageNet (source) and then tested on four different versions of ImageNet (targets) that exhibit some kind of domain shift. Average performance improvements over the baselines are shown in **green**, and average performance deterioration in **red**.

		Source	Target				
Backbone	Method	ImageNet	-V2	-S	-A	-R	*Average*
RN50	CLIP (student)	58.20	51.50	33.30	21.70	56.00	40.63
	CoOp	62.40	55.17	33.70	23.13	56.20	42.05
	CoOp + KDPL	**62.73**	**55.37**	**35.20**	**23.27**	**57.77**	**42.90**
	CoCoOp	**63.07**	55.53	34.77	**23.73**	**59.47**	43.38
	CoCoOp + KDPL	62.70	**55.60**	**35.30**	23.43	57.90	**43.06**
ViT-B/32	CLIP (student)	62.00	54.70	40.80	29.60	66.00	47.78
	CoOp	**66.33**	**58.30**	41.40	31.47	65.87	49.26
	CoOp + KDPL	65.97	58.10	**42.50**	**31.63**	**67.37**	**49.90**
	VPT	64.97	56.73	41.27	27.00	66.50	47.88
	VPT + KDPL	**65.10**	**57.37**	**41.67**	**27.77**	**67.47**	**48.57**
	MaPLe	**66.80**	**58.53**	42.23	**30.13**	66.40	49.32
	MaPLe + KDPL	66.50	58.47	**42.77**	29.87	**67.70**	**49.70**
	PromptSRC	**66.33**	**58.70**	42.97	**32.07**	**68.93**	50.67
	PromptSRC + KDPL	66.27	58.60	**43.07**	31.70	68.83	**50.55**
ViT-H/14	CLIP (teacher)	82.80	76.60	71.10	71.10	91.30	77.53

Experiment

- Zero Shot Classification(Cross-dataset Transfer)
  - Source Datasets: imagenet1k
  - Target Datasets: OxfordPets, Flowers102 등 총 10개의 Datasets

Table 2: Cross-dataset Transfer. Comparison between the baselines and the proposed unsupervised KDPL variants (highlighted in cyan). The prompt is learned on ImageNet and then tested on the ten different target datasets. Average performance improvements over the baselines are indicated in **green**.

		Target										
Backbone	Method	*OxfordPets*	*Flowers102*	*FGVCAircraft*	*DTD*	*EuroSAT*	*StanfordCars*	*Food101*	*SUN397*	*Caltech101*	*UCF101*	*Average*
RN50	CLIP (student)	83.70	61.00	15.60	40.00	24.20	55.60	75.20	58.50	86.00	58.30	55.81
	CoOp	84.60	61.63	13.77	36.83	22.33	54.20	75.03	59.00	87.67	57.73	55.28
	CoOp + KDPL	**87.33**	**62.83**	**14.97**	**39.27**	**30.17**	**56.60**	**77.27**	**60.30**	**88.57**	**58.63**	**57.59**
	CoCoOp	86.93	63.47	**16.27**	40.53	27.20	55.47	77.73	61.40	88.07	**60.43**	57.75
	CoCoOp + KDPL	**87.33**	**64.97**	15.97	**41.27**	**29.17**	**56.83**	**78.00**	**61.50**	**88.50**	59.53	**58.31**
ViT-B/32	CLIP (student)	85.00	64.30	18.20	42.80	38.10	60.40	79.10	62.00	91.10	60.70	60.17
	CoOp	**87.13**	62.20	12.13	**40.57**	36.97	57.83	80.13	62.83	91.13	**61.80**	59.27
	CoOp + KDPL	**87.13**	**62.23**	**15.57**	40.40	**40.93**	**58.83**	**80.57**	**64.23**	**92.77**	61.63	**60.43**
	VPT	**87.43**	**64.77**	**19.53**	**44.20**	28.47	57.60	78.27	**63.73**	91.40	62.50	59.79
	VPT + KDPL	**87.43**	64.43	17.77	42.93	**30.80**	**57.97**	**78.47**	63.63	**92.00**	**62.77**	**59.82**
	MaPLe	**88.37**	**66.43**	**18.47**	42.03	**37.77**	**60.00**	80.10	64.43	91.33	62.50	61.14
	MaPLe + KDPL	88.27	64.77	17.33	**43.13**	37.13	59.93	**80.47**	**65.43**	**93.43**	**62.70**	**61.26**
	PromptSRC	88.13	**65.33**	17.17	**43.90**	40.50	60.17	**81.23**	65.40	92.37	**63.63**	61.78
	PromptSRC + KDPL	**88.17**	65.10	**18.23**	43.27	**44.30**	**60.43**	80.90	**65.67**	**93.63**	63.40	**62.31**
ViT-H/14	CLIP (teacher)	94.80	89.40	63.20	66.80	63.30	95.60	93.60	76.40	97.90	76.40	81.74

Experiment

- Zero Shot Classification(Generalization to Unseen Classes)
  - Datasets: Imagenet, OxfordPets, Flowers102 등 총 11개의 Datasets

Table 3: Generalization to Unseen Classes. Comparison between the baselines and the proposed unsupervised KDPL variants (highlighted in cyan). In this setting the prompt is tested on classes never observed during training. Average performance improvements over the baselines are indicated in **green**.

		Unseen Classes											
Backbone	Method	*ImageNet*	*OxfordPets*	*Flowers102*	*FGVCAircraft*	*DTD*	*EuroSAT*	*StanfordCars*	*Food101*	*SUN397*	*Caltech101*	*UCF101*	*Average*
RN50	CLIP (student)	60.10	93.70	71.30	24.80	53.90	43.70	66.60	82.20	70.10	90.50	67.80	65.88
	CoOp	60.00	91.63	58.30	**22.17**	**40.93**	42.13	**59.87**	82.03	66.77	87.53	57.63	60.82
	CoOp + KDPL	**60.53**	**94.97**	**63.40**	21.27	38.70	**62.07**	56.63	**83.07**	**69.83**	**88.75**	**62.63**	**63.80**
	CoCoOp	**63.07**	94.93	67.47	**24.87**	45.40	35.10	**64.73**	84.50	72.87	**90.73**	64.67	64.39
	CoCoOp + KDPL	62.43	**95.10**	**69.17**	24.57	**48.00**	**55.93**	63.70	**85.67**	**73.33**	90.53	**66.40**	**66.80**
ViT-B/32	CLIP (student)	64.00	96.30	72.30	28.30	53.90	61.50	69.80	85.70	73.10	94.00	71.60	70.05
	CoOp	64.53	93.37	**59.77**	**23.03**	45.57	49.13	61.73	85.50	69.37	92.13	65.60	64.52
	CoOp + KDPL	**64.90**	**94.00**	56.17	22.00	**47.33**	**62.07**	**62.37**	**87.23**	**74.17**	**93.43**	**66.20**	**66.35**
	VPT	64.60	94.83	66.03	29.70	**50.17**	50.03	**69.90**	85.83	75.13	92.57	69.47	68.02
	VPT + KDPL	**65.20**	**95.20**	**67.87**	**30.40**	47.53	**54.97**	69.80	**86.33**	**75.17**	**92.80**	**71.07**	**68.76**
	MaPLe	66.70	**96.87**	**69.77**	20.43	**53.50**	66.70	68.03	87.47	76.57	92.27	69.40	69.79
	MaPLe + KDPL	**66.73**	**96.87**	68.23	**27.37**	50.33	**75.70**	**68.40**	**87.83**	**76.83**	**93.67**	**73.53**	**71.41**
	PromptSRC	65.93	96.30	**71.37**	23.67	**54.43**	**61.97**	**70.00**	87.13	76.77	**94.70**	72.63	70.45
	PromptSRC + KDPL	**66.33**	**96.47**	68.83	**28.43**	50.87	74.17	68.50	**87.63**	**77.33**	94.47	**74.73**	**71.61**
ViT-H/14	CLIP (teacher)	84.10	99.20	85.60	64.30	75.00	82.60	98.20	96.30	85.10	97.30	85.20	86.63

Conclusion

- Pros
  - Class Agnostic 상황에서도 학습 가능한 Label free Method
  - Prompt Learning 기법에 적용 가능한 범용적인 Plugin KD Method
  - 2 Stage를 필요로 하지 않음

- Cons
  - 성능 향상이 미미
    - Teacher의 성능이 효과적으로 Distillation되지 못한 것으로 보임

Reference

- [1] Li, Zheng, et al. "Promptkd: Unsupervised prompt distillation for vision-language models." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.
- [2] Saha, Oindrila, Grant Van Horn, and Subhransu Maji. "Improved Zero-Shot Classification by Adapting VLMs with Text Descriptions." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.
- [3] Zhou, Kaiyang, et al. "Learning to prompt for vision-language models." *International Journal of Computer Vision* 130.9 (2022): 2337-2348.
- [4] Zhou, Kaiyang, et al. "Conditional prompt learning for vision-language models." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.
- [5] Bahng, Hyojin, et al. "Exploring visual prompts for adapting large-scale models." *arXiv preprint arXiv:2203.17274* (2022).
- [6] Jia, Menglin, et al. "Visual prompt tuning." *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, 2022.
- [7] Khattak, Muhammad Uzair, et al. "Maple: Multi-modal prompt learning." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023.
- [8] Khattak, Muhammad Uzair, et al. "Self-regulating prompts: Foundational model adaptation without forgetting." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023.

감사합니다.

Surplus Material

- CLIP 모델 별 Parameter 및 Detail

	Training Data	**Patch Size**	**Parameter**	**#Tokens**	**Embedding Dimension**
ViT-H/14 layer32	LAION-2B	14x14	632M	249+1	**1024**
ViT-L/14	LAION-2B	14x14	304M	249+1	768
ViT-B/16	DataComp-1B	16x16	86.2M	224+1	512
ViT-B/32	DataComp-1B	32x32	86.2M	49+1	512
ViT-T/16		16x16	5.6M	224+1	384

Surplus Material

- Dataset Statistics

Table 5: Datasets statistics.

Dataset	Classes	Train	Val	Test
ImageNet [12]	1,000	1.28M	N/A	50,000
Caltech101 [20]	100	4,128	1,649	2,465
OxfordPets [43]	37	2,944	736	3,669
StanfordCars [32]	196	6,509	1,635	8,041
Flowers102 [41]	102	4,093	1,633	2,463
Food101 [6]	101	50,500	20,300	30,300
FGVCAircraft [39]	100	3,334	3,333	3,333
SUN397 [56]	397	15,880	3,970	19,850
DTD [11]	47	2,820	1,128	1,692
EuroSAT [23]	10	13,500	5,400	8,100
UCF101 [53]	101	7,639	1,898	3,783
ImageNetV2 [45]	1,000	N/A	N/A	10,000
ImageNet-S [54]	1,000	N/A	N/A	50,889
ImageNet-A [25]	200	N/A	N/A	7,500
ImageNet-R [24]	200	N/A	N/A	30,000