2025 동계 세미나

Prompt Tuning for Image Classification with Linguistic Knowledge Augmentation

Sogang University
Vision & Display Systems Lab, Dept. of Electronic Engineering

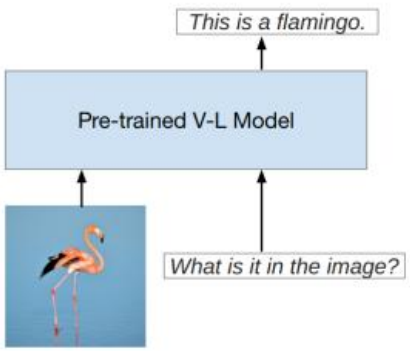
Presented By
장순원

Outline

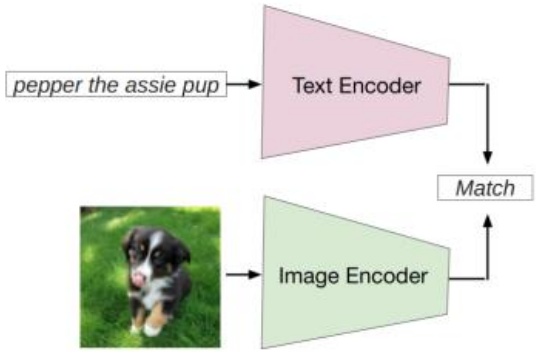
- Introduction
  - Text prompt tuning for VLM
- LLaMP: Large Language Models are Good Prompt Learners for Low-Shot Image Classification
  - CVPR 2024
- AWT: Transferring Vision-Language Models via Augmentation, Weighting, and Transportation
  - NeurIPS 2024

Introduction

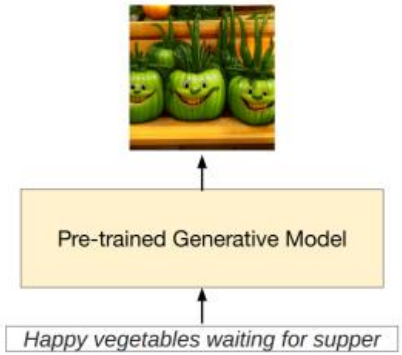
- Vision-Language Model Foundation Models
  - Multimodal-to-Text Generation
  - Text-to-Image Generation
  - Image-Text Matching
    - Image-text pair를 같은 embedding space에 mapping시킬 수 있도록 사전학습 수행
    - Web scale의 대규모 데이터셋을 통해 학습되어 image encoder 자체도 좋은 성능을 보임
    - Zero-shot classification과 우수한 few-shot 성능으로 인해 다양한 domain에서 활용됨



(a) Multimodal-to-Text Generation



(b) Image-Text Matching

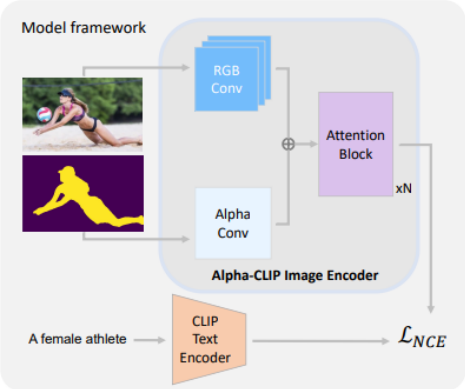


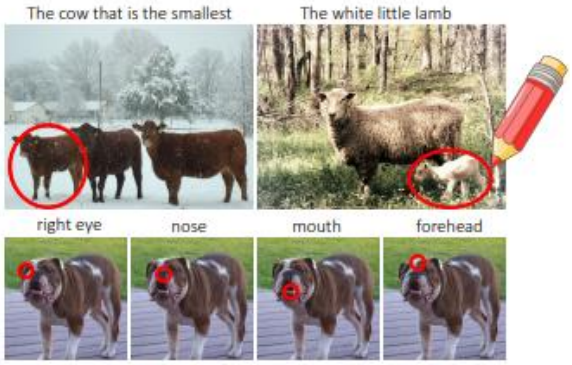
(c) Text-to-Image Generation

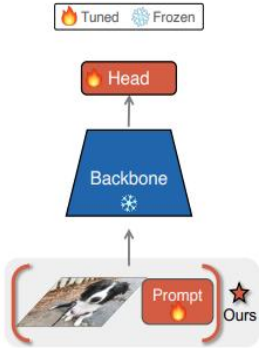
1) Sun, Zeyi, et al. "Alpha-clip: A clip model focusing on wherever you want." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.
2) Jia, Menglin, et al. "Visual prompt tuning." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2022.
3) Shtedritski, Aleksandar, Christian Rupprecht, and Andrea Vedaldi. "What does CLIP know about a red circle? Visual prompt engineering for VLMs. In 2023 IEEE." CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE. 2023.

Introduction

- Text prompt tuning
  - Prompt
    - 특정 task에 대한 추가 정보나 지시를 목적으로 모델에 입력되는 추가 input
    - Hard prompt
      - 모델에 명시적으로 입력되는 고정된 형태의 추가 정보
      - Text, Box, Segmentation mask, Synthetic image, etc.
    - Soft prompt
      - Additional learnable input
    - VLM에서는 text prompt가 주요한 역할을 하며 classification을 보조하게 됨

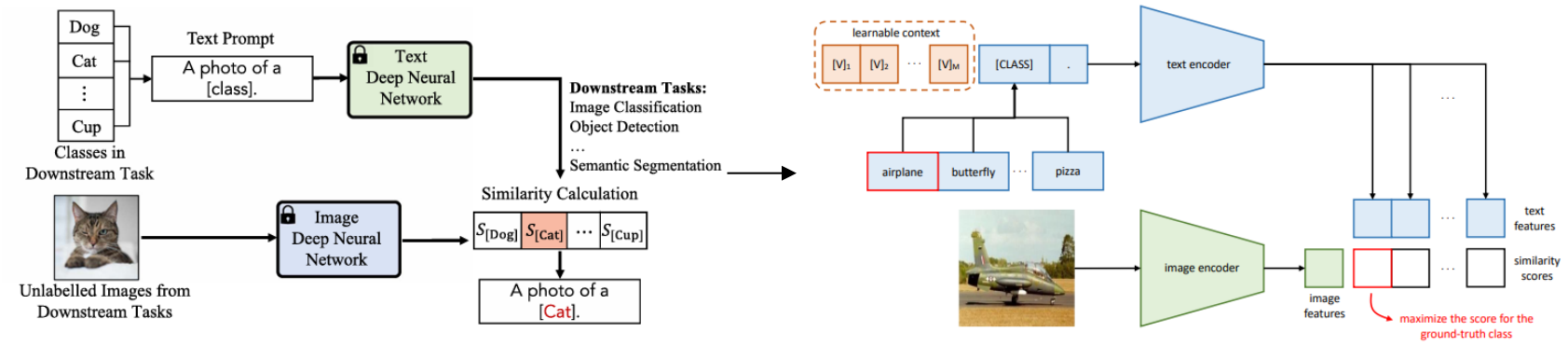






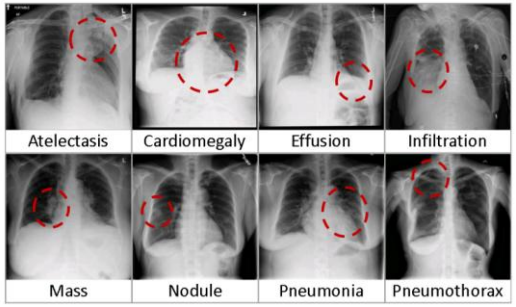
Introduction

- Text prompt tuning
  - Class를 대표하는 정확한 text prompt의 설계를 통해 성능을 크게 개선할 수 있음
  - 초기 논문에서는 “a photo of a {class}”와 같은 prompt를 적용함
    - Prompt engineering에 매우 민감하며 최적화가 어려움
  - Context Optimization (CoOp)
    - Class name을 제외한 prompt의 나머지 부분을 learnable parameter로 설정하여 학습
    - Few-shot learning에서 높은 성능을 보여줌
    - In-domain에서 학습된 prompt를 out-of-domain 데이터에 적용 가능



Introduction

- Text prompt tuning
  - CoOp에 대한 문제 제기
    - 모든 class에 대해 동일한 prompt 사용
    - Class name으로 얻을 수 있는 정보는 한정적임
  - Knowledge augmentation
    - 특수한 domain(ex. Medical, Aircraft)에서 CLIP 활용의 한계가 존재
      - Class에 대한 정보를 text description을 통해 제공하여 부족한 image 정보 보완
    - Text modality의 특성을 활용하고자 함
      - High semantic 정보 및 지시 전달에 용이, Language Model과의 연결성



< ChestX-ray8 dataset >

< FGVCAircraft dataset >

- LLaMP: Large Language Models are Good Prompt Learners for Low-Shot Image Classification (CVPR 2024)

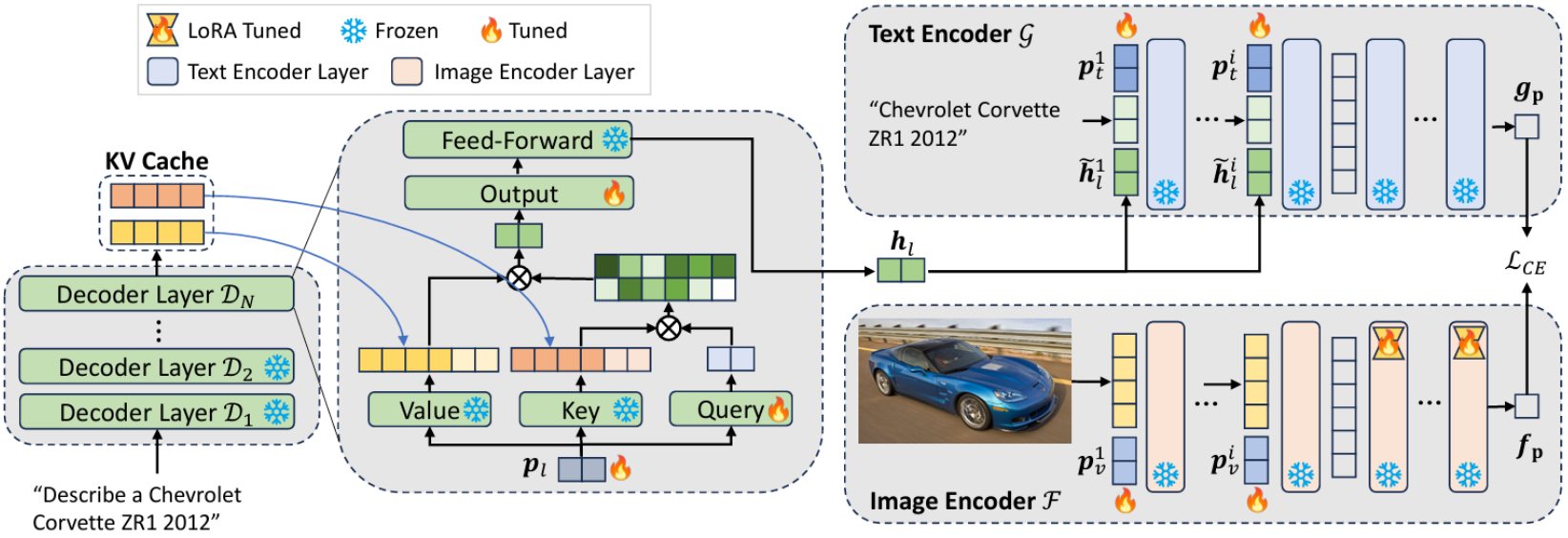
1) Zheng, Zhaoheng, et al. "Large Language Models are Good Prompt Learners for Low-Shot Image Classification." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.

LLaMP[1)]

- Introduction
  - LLM의 knowledge를 통해 CLIP에 입력되는 text prompt의 정보 증강
    - LLM의 사전적 지식을 활용하여, 클래스 이름만으로는 부족할 수 있는 세부 정보와 특징을 추가로 제공
  - LLM을 CLIP에 적합한 class-specific prompt generator로 학습
    - Prompt 자체가 아닌 prompt를 생성할 수 있는 모델을 학습
  - LLM을 vision task와 연결시키기 위한 효율적인 fine-tuning method 실험 및 적용

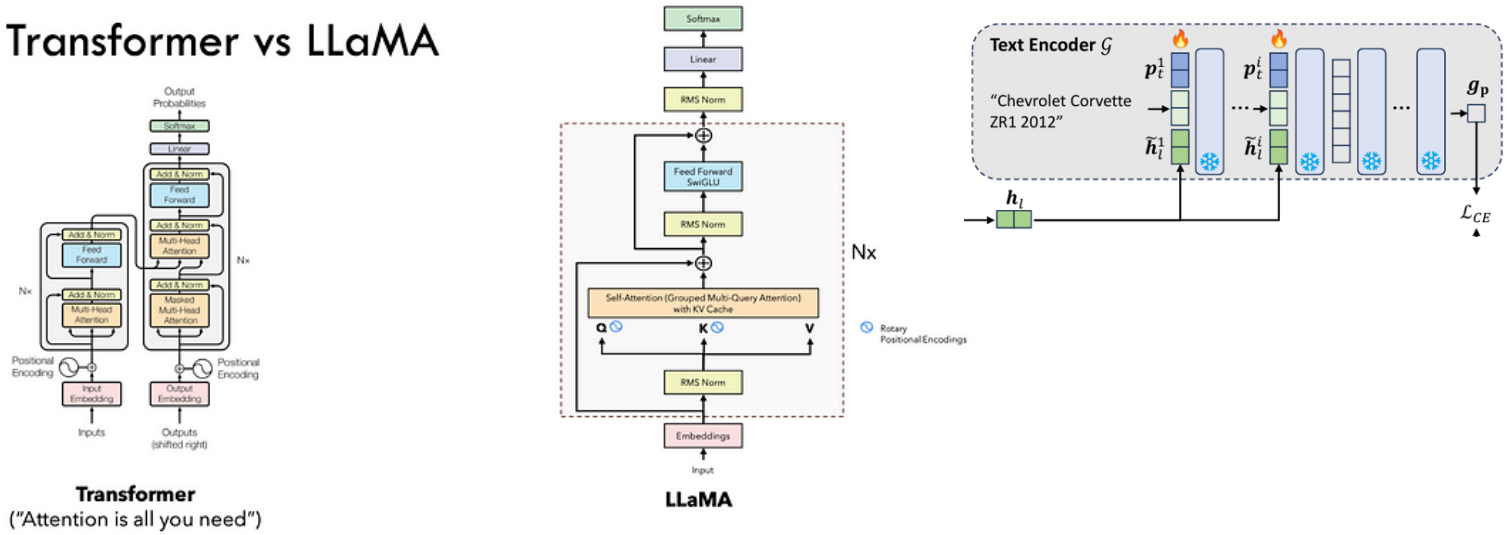
LLaMP[1)]

- Overall framework
  - LLM을 통한 class-specific text prompt generation
    - CLIP Text Encoder input
      - Learnable text prompt p_t^i, Template + Class name, Class-specific prompt $\tilde{h}_l^i$
  - LLM과 CLIP의 다양한 fine-tuning method 적용



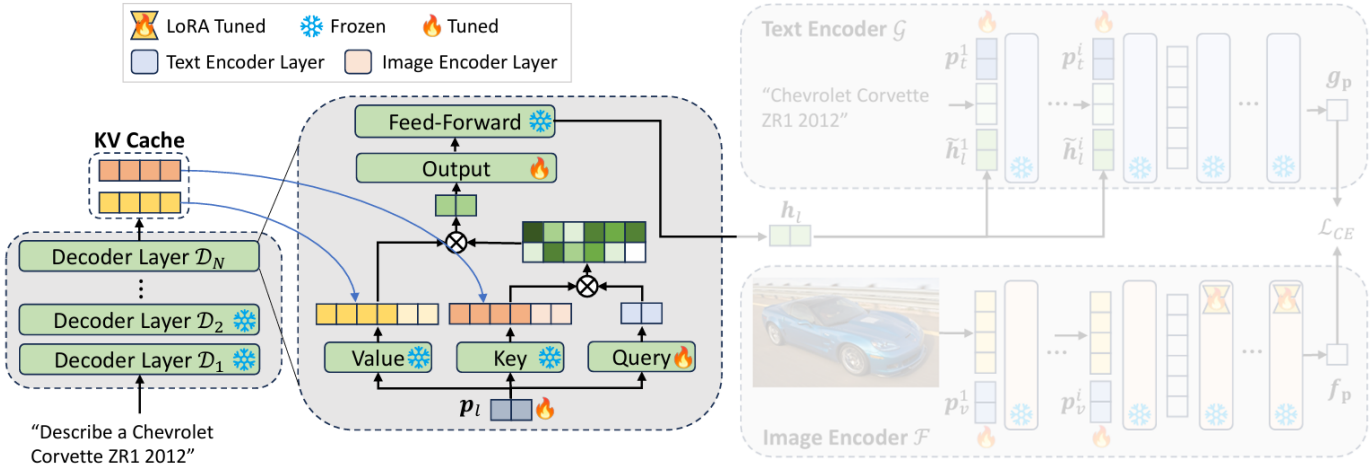
LLaMP[1)]

- Class-specific prompt ($\tilde{h}_l^i$) generate via LLM
  - LLaMa2-7B를 사용
    - Decoder only 구조로 decoder의 self-attention layer에서 preceding token 정보만 필요
  - K개의 learnable prompt를 input에 포함시켜 해당 위치의 output sequence h_l 획득
    - CLIP의 training 과정에 포함시키기 용이하도록 한 번의 forward path를 통해 output을 얻을 수 있도록 설정
    - h_l은 linear layer를 거쳐 CLIP text encoder의 각 layer에 입력됨



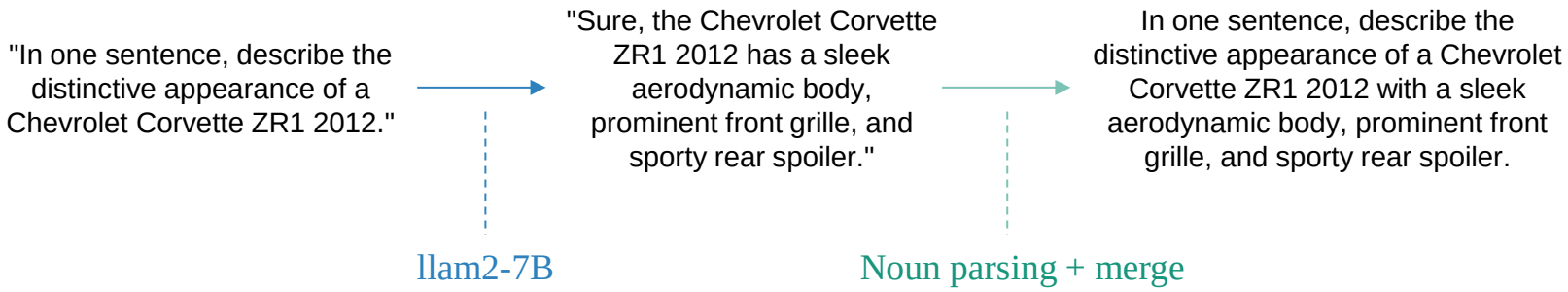
LLaMP[1)]

- Class-specific prompt ($\tilde{h}_l^i$) generate via LLM
  - LLM Knowledge Cache
    - Prompt 자체의 parameter는 적지만 LLM의 forward-backward path가 모두 수행되어야 하기 때문에 여전히 적지 않은 training cost를 가지고 있음
    - 마지막 layer에만 prompt를 추가하면 마지막 layer와 LLM query의 KV Cache만을 통해 효율적인 fine tuning 수행 가능



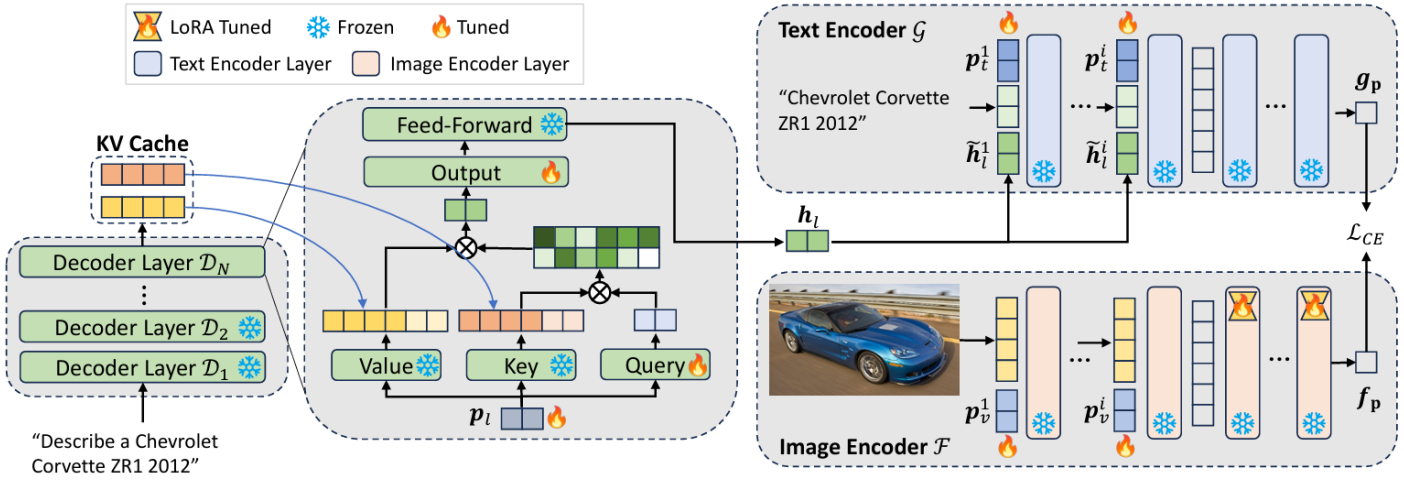
LLaMP[1)]

- Text augmentation
  - LLM input query
    - Initial prompt augmentation
      - NLP engine을 통해 LLM의 출력에서 noun phrase를 가져와 initial prompt와 결합
    - Diverse initial prompt
      - "In one sentence, describe the distinctive appearance of {class}" 기반
  - Template for CLIP text encoder
    - "a photo of {} with {NP}"
  - 학습 시에는 여러 종류의 template 중 하나를 iteration마다 랜덤하게 선택



LLaMP[1)]

- Fine-tuning target
  - Learnable prompt for each model (p_l, p_t^i, p_v^i)
  - Query, Output Projection layer in last layer of LLM decoder
  - Additional LoRA fine-tuning for image encoder (Query, Value projection layer)
    - Prompt x6 + LoRA x 6



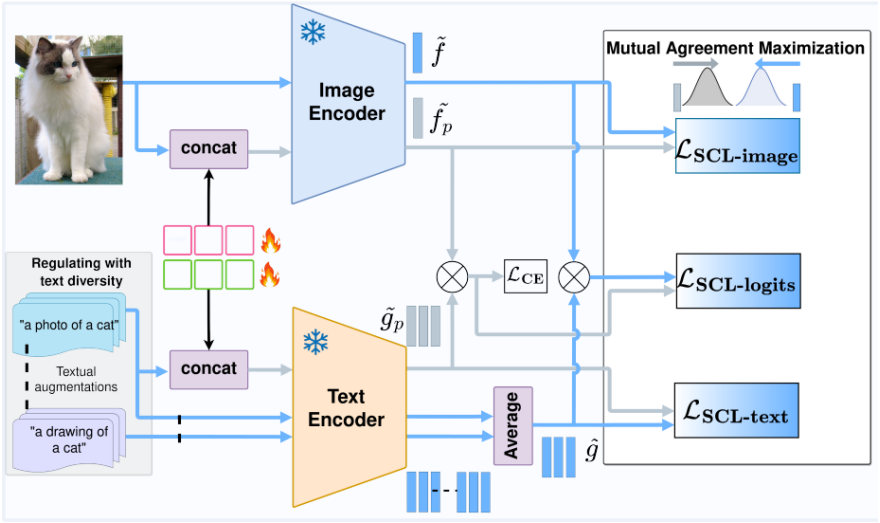
LLaMP[1)]

- Training objective
  - Cross-entropy loss
  - Self-regularization loss from PromptSRC[2)]
    - Feature-level, logit-level에서 pretrained CLIP과의 distillation loss 적용
      - Feature-level은 L1-distance, logit-level은 KL divergence를 통해 distance measure
    - Downstream task에 맞게 fine-tuning 되면서도 CLIP의 일반화 성능을 유지할 수 있도록 하기 위한 regularization 역할

$$\mathcal{L}_{CE} = -\frac{1}{\mathcal{N}} \sum_i \log \frac{\exp\left(\boldsymbol{f}_{\boldsymbol{p}}^i \cdot \boldsymbol{g}_{\boldsymbol{p}}^i / \tau\right)}{\sum \exp\left(\boldsymbol{f}_{\boldsymbol{p}}^i \cdot \boldsymbol{g}_{\boldsymbol{p}}^j / \tau\right)}$$

$$\mathcal{L}_{l1} = \frac{1}{\mathcal{N}} \sum_i \lambda_v |\boldsymbol{f}_{\boldsymbol{p}}^i - \hat{\boldsymbol{f}}^i| + \frac{1}{C} \sum_i \lambda_t |\boldsymbol{g}_{\boldsymbol{p}}^i - \hat{\boldsymbol{g}}^i|$$

$$\mathcal{L}_{dist} = \lambda_{dist} D_{KL}(\boldsymbol{f_p} \cdot \boldsymbol{g_p}, \hat{\boldsymbol{f}} \cdot \hat{\boldsymbol{g}})$$



< PromptSRC[2)] >

LLaMP[1)]

- Experiment
  - Generalization to unseen classes
    - Evaluation strategy
      - Base class에서 학습 후 Novel class에서의 accuracy를 평가하여 일반화 성능을 확인
      - Harmonic mean을 통해 ID, OOD에서의 종합적인 성능을 측정
    - 11 dataset (ImageNet, Caltech101, OxfordPets, StanfordCars, Flowers102, Food101, FGVCAircraft, SUN397, UCF101, DTD, EuroSAT)

Method	*Average*			ImageNet [8]			Caltech101 [10]			OxfordPets [29]		
	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM	Base	Novel	HM
CLIP [32]	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 [49]	82.69	63.22	71.66	76.47	67.88	71.92	98.00	89.81	93.73	93.67	95.29	94.47
CoCoOp [48]	80.47	71.69	75.83	75.98	70.43	73.10	97.96	93.81	95.84	95.20	97.69	96.43
KAPT* [17]	78.41	70.52	74.26	71.10	65.20	68.02	97.10	93.53	95.28	93.13	96.53	94.80
ProDA [24]	81.56	72.30	76.65	76.66	70.54	73.47	97.74	94.36	96.02	95.43	97.76	96.58
MaPLe [18]	82.28	75.14	78.55	75.40	70.32	72.72	98.27	93.23	95.68	95.43	**97.83**	96.62
RPO [23]	81.13	75.00	77.78	76.60	**71.57**	74.00	97.97	94.37	96.03	94.63	97.50	96.05
PSRC [19]	84.26	76.10	79.97	77.60	70.73	74.01	98.10	94.03	96.02	95.33	97.30	96.30
LLaMP	**85.16**	**77.71**	**81.27**	**77.99**	71.27	**74.48**	**98.45**	**95.85**	**97.13**	**96.31**	97.74	**97.02**
Δ w.r.t. PSRC	+0.90	+1.61	+1.30	+0.39	+0.54	+0.47	+0.35	+1.82	+1.11	+0.98	+0.44	+0.72

LLaMP[1)]

- Experiment
  - Few-shot classification

	16-Shot Classification											
	Average	ImageNet [8]	Caltech [10]	Pets [29]	Cars [21]	Flowers [26]	Food [3]	Aircraft [25]	SUN397 [45]	DTD [7]	EuroSAT [11]	UCF101 [37]
CLIP [32]	78.79 (65.02)	67.31	95.43	85.34	80.44	97.37	82.90	45.36	73.28	69.96	87.21	82.11
CoOp [49]	79.89 (73.82)	71.87	95.57	91.87	83.07	97.07	84.20	43.40	74.67	69.87	84.93	82.23
CoCoOp [48]	74.90 (70.70)	70.83	95.16	93.34	71.57	87.84	87.25	31.21	72.15	63.04	73.32	78.14
MaPLe [18]	81.79 (75.58)	72.33	96.00	92.83	83.57	97.00	85.33	48.40	75.53	71.33	92.33	85.03
PSRC [19]	82.87 (77.90)	73.17	96.07	93.67	83.83	97.60	87.50	50.83	**77.23**	72.73	**92.43**	86.47
LLaMP	**83.81 (78.50)**	**73.49**	**97.08**	**94.21**	**86.07**	**98.06**	**87.62**	**56.07**	77.02	**74.17**	91.31	**86.84**

  - Ablation study

Method	Base	Novel	HM
LLM Only	81.74	35.82	49.81
LLaMP	**85.16**	**77.71**	**81.27**

< Align without CLIP >

Scheme	Base	Novel	HM
Prompt × 9	84.67	77.28	80.81
LoRA × 12	84.89	77.27	80.90
Prompt ×6 + LoRA ×6	**85.16**	**77.71**	**81.27**

< Image encoder training scheme >

LLaMP[1)]

- Conclusion
  - Contributions
    - CLIP을 bridge로 LLM의 지식을 image classification model과 연결 가능함을 보임
    - Zero-shot generalization과 few-shot learning에서 기존 모델을 능가하는 성능을 달성함
    - Train data를 통해서 class-specific text prompt를 추출하는 법을 학습함
      - Class마다 다른 prompt를 적용하면서도 일반화 성능을 높일 수 있음
  - Limitations
    - LLM에 입력되는 초기 프롬프트 설정 및 증강 과정이 수동으로 이루어져 최적화가 어려움
    - LLM의 지식과 성능에 의존
    - Prompt tuning시 실제 sample image 정보는 사용하지 않음

- AWT: Transferring Vision-Language Models via Augmentation, Weighting, and Transportation (NeurIPS 2024)

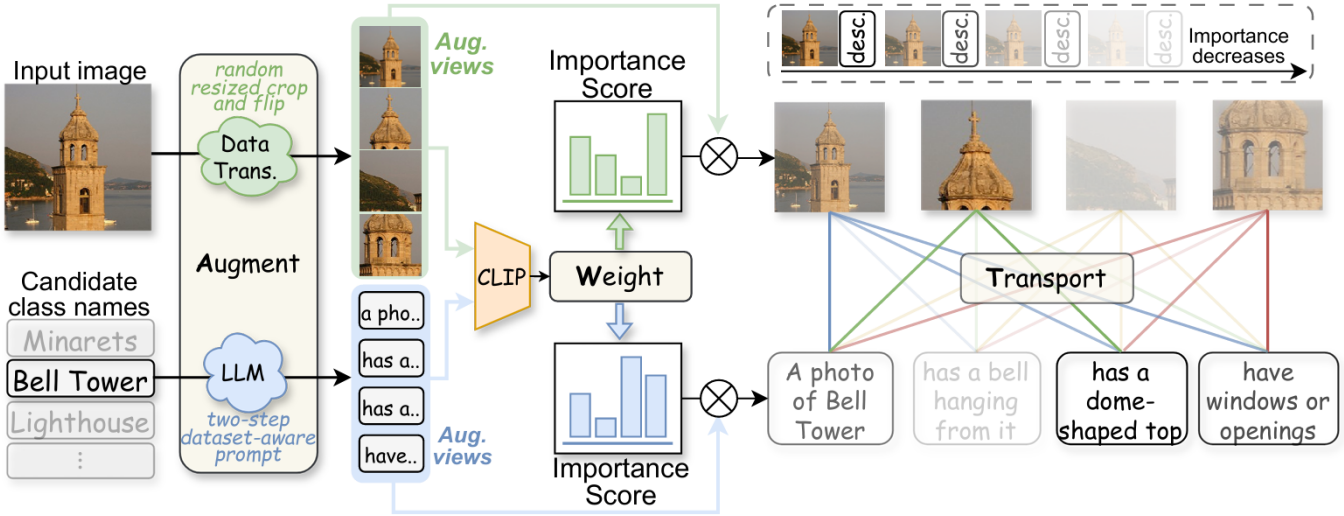
1) Zhu, Yuhan, et al. "Awt: Transferring vision-language models via augmentation, weighting, and transportation." *arXiv preprint arXiv:2407.04603* (2024).

AWT[1)]

- Introduction
  - Augmented prompt를 효과적으로 활용하는 방법 탐색
    - Class에 대한 descriptions를 특정 region과 연결시키며 추가 training 없이 높은 성능 향상을 보여줌
    - Classification에 필요한 specific interest region에 집중하는 능력이 부족한 CLIP의 단점 보완
  - Hard prompt를 사용하여 모델 해석 및 추가 활용에 용이
    - 단순 embedding vector 형태가 아닌 해석 가능한 인간 언어를 사용하여 모델 동작의 해석 및 수정이 용이함

AWT[1)]

- Overall framework
  - 목적에 따라 Augment – Weight – Transport의 3 stage로 구성
    - Augment를 통해 image, text views를 생성하고 weight-transport를 통해 classification score 계산
  - LLM은 class name과 domain 정보가 주어졌을 때 description을 자동으로 생성할 수 있는 pipeline 구축을 위해 사용됨



AWT[1)]

- Augment
  - Visual augmentation
    - Random resized crop, Radom flip을 통해 다양한 view 생성
  - Text augmentation
    - Class를 설명하는 다양한 description 작성
    - Two-step, dataset-aware prompt strategy for LLM
      - Initial prompt: "Generate questions to classify images from a dataset, which {dataset description}"
      - Final prompt: "Describe a {class}" + initial prompt로부터 얻은 결과
  - Raw image, text input을 포함하여 각각 N+1, M+1의 view를 사용

They tend to have sturdy, straight **legs** and a curled **tail**.
They have a **thick, fluffy white coat**.
Its **ears are erect and triangular**, covered in fur.
The **nose is black**, contrasting with its **white fur**.

AWT[1)]

- Weight
  - Entropy-based weighting
  - 가장 중요한 view는 classification confidence가 가장 높은 view로 가정
    - 균일한 분포일수록 높은 값을 가지는 entropy를 통해 confidence를 측정

      $H(x) = -\sum_x p(x) log p(x)$
    - 'Raw image'와 'class별 text embedding 평균'을 기준으로 classification probability 계산
    - Probability에 negative entropy에 softmax를 적용하여 가중치 계산

$$\mathbf{a}_n = \frac{\exp\left(-H_n(t)/\gamma_v\right)}{\sum_{j=1}^{N+1}\exp\left(-H_j(t)/\gamma_v\right)} \qquad \mathbf{b}_m^i = \frac{\exp\left(-H_m^i(t)/\gamma_t\right)}{\sum_{k=1}^{M+1}\exp\left(-H_k^i(t)/\gamma_t\right)}$$

	Prob.
Horse	0.1
Cat	0.7
Dog	0.2

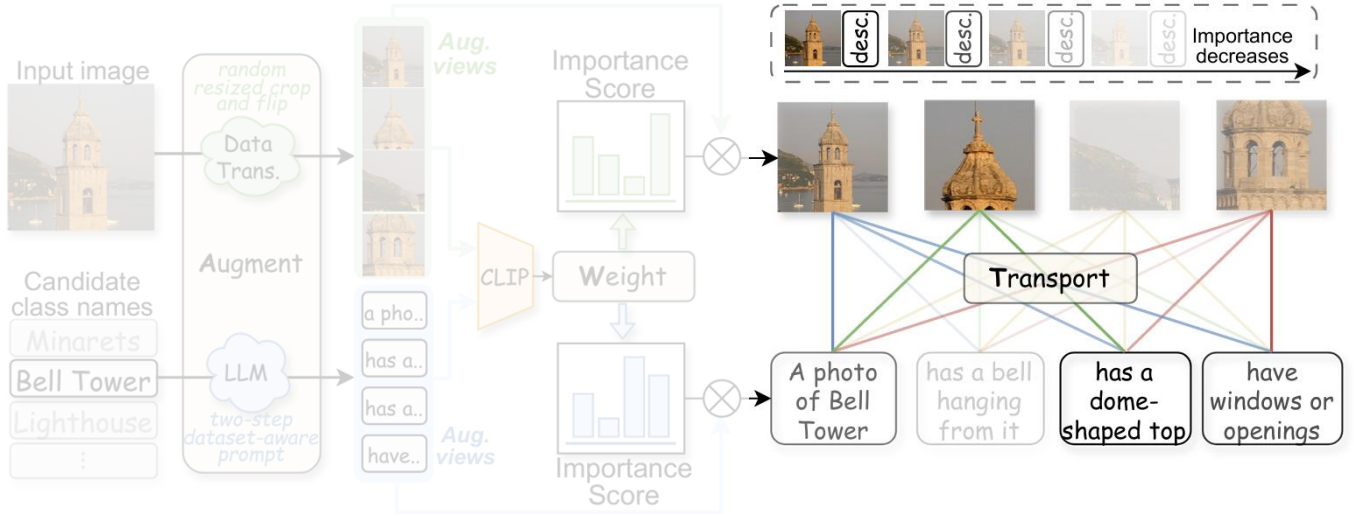
< high confidence image view >

	Prob.
Horse	0.3
"triangular ears"	0.4
Dog	0.3

< low confidence text view >

AWT[1)]

- Transport
  - Classification score 계산 목적
    - 하나의 이미지, 하나의 클래스에 대해 여러 embedding이 존재하기 때문에 둘의 거리를 측정하기 위한 새로운 방식이 필요
      - Embeddings의 평균끼리 거리를 비교하는 경우는 cross-modal correlation을 제대로 반영할 수 없음
      - 특히 “where specific textual descriptions correlate directly with certain image crops”의 경우에 문제가 발생할 수 있음



AWT[1)]

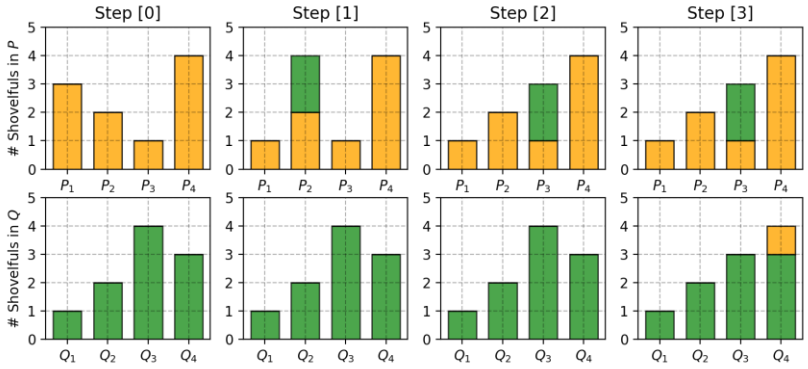
- Transport
  - Wasserstein distance (Earth mover's distance)
    - 한 분포에서 다른 분포로 물질을 이동시키는 데 드는 최소 비용
      - 거의 동일한 위치에 mapping된 두 embedding은 거리에 거의 영향을 주지 않음

$$W(p,q) = \inf_{\gamma \in \Pi(P,Q)} \mathbb{E}_{(x,y)\sim\gamma}[\|x-y\|]$$

P와 Q의 가능한 모든 joint distribution의 집합

    - Weight stage에서 구한 가중치를 통해 embedding space에서의 distribution function을 정의
      - Embedding간의 distance measure는 $1 - cosine\ similarity$ 사용

$$\alpha = \sum_{n=1}^{N+1} \mathbf{a}_n \delta_{\boldsymbol{I}^n} \quad \text{and} \quad \left\{ \beta^i = \sum_{m=1}^{M+1} \mathbf{b}_m^i \delta_{\boldsymbol{e}_i^m} \right\}_{i=1}^{C}$$



AWT[1)]

- Experiment
  - Zero-shot classification

	Train	IN-1k[89]	Flowers[90]	DTD[91]	Pets[92]	Cars[93]	UCF[94]	Cal101[95]	Food[96]	SUN[97]	Aircraft[98]	SAT[99]	Birds[100]	Cal256[101]	CUB[102]	Avg.(14)
CLIP [1]	✗	66.74	67.44	44.27	88.25	65.48	65.13	93.35	83.65	62.59	23.67	42.01	42.80	82.50	54.90	63.06
CoOp [9]	✓	71.51	68.71	41.92	89.14	64.51	66.55	93.70	85.30	64.15	18.47	46.39	41.43	82.91	53.18	63.42
CoCoOp [13]	✓	71.02	71.88	45.73	90.14	65.32	68.21	94.43	86.06	67.36	22.94	45.37	43.75	85.39	56.09	65.26
MaPLe [15]	✓	70.72	72.23	46.49	90.49	65.57	68.69	93.53	86.20	67.01	24.74	48.06	44.06	85.58	57.18	65.75
PLOT++ [83]	✓	72.48	69.10	38.42	90.49	61.20	68.94	91.32	86.07	61.59	24.84	49.90	36.37	84.30	48.58	63.11
POMP [22]	✓	70.16	72.72	44.44	89.05	66.70	68.44	94.65	86.28	67.27	25.47	52.65	43.94	86.76	56.92	66.10
ProVP-Ref [32]	✓	71.14	71.62	45.97	91.58	65.29	67.72	93.79	86.17	66.29	24.51	51.95	–	–	–	66.91
TPT [49]	✓	68.98	68.98	47.75	87.79	66.87	68.04	94.16	84.67	65.50	24.78	42.44	44.56	85.71	56.97	64.80
DiffTPT [50]	✓	70.30	70.10	47.00	88.22	67.01	68.22	92.49	**87.23**	65.74	25.60	43.13	–	–	–	65.91
PromptAlign [52]	✓	71.44	72.39	47.24	90.76	68.50	69.47	94.01	86.65	67.54	24.80	47.86	45.28	86.05	57.90	66.42
Self-TPT-v [10]	✓	72.96	71.79	49.35	91.26	68.81	69.50	94.71	85.41	68.18	27.57	51.91	–	–	–	68.31
CuPL [63]	✗	69.62	71.30	44.56	89.13	65.29	66.83	92.98	86.11	62.59	24.90	47.84	41.24	86.30	56.28	64.64
VisDesc [61]	✗	68.55	70.85	44.98	88.85	64.08	67.12	94.60	85.05	67.99	24.30	54.84	43.64	87.16	56.59	65.61
WaffleCLIP [62]	✗	68.81	72.35	45.21	89.95	63.57	67.19	94.02	86.68	67.23	25.39	55.07	43.92	87.04	57.17	65.97
SuS-X-SD [65]	✗	69.88	73.81	54.55	90.57	66.13	66.59	93.96	86.08	67.73	28.68	57.49	45.53	87.45	57.11	67.54
AWT	✗	**71.32**	**75.07**	**55.56**	**92.53**	**69.93**	**72.51**	**95.54**	85.54	**70.58**	**29.22**	**58.61**	**48.75**	**88.84**	**60.20**	**69.59**

AWT[1)]

- Experiment
  - Ablation study
    - 50개 이상의 view에서는 성능 변화가 거의 존재하지 않음

(a) **Main component analysis**: Augment(A), Weight(W) and Transport(T).

Step	OOD	Avg.(14)
Raw inputs	57.15	63.06
A(Img.)	56.77	63.40
A(Name)	59.76	67.36
A(Img.+Name)	59.92	67.13
A+T	60.48	67.80
A+W+T	**64.43**	**69.59**

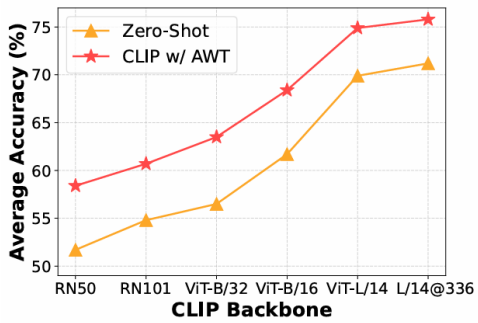
(b) Number of **augmented image views** N.

N	OOD	Avg.(14)
2	61.40	68.57
5	62.59	68.98
10	63.31	69.16
25	64.13	69.33
50	64.43	**69.59**
100	**64.54**	69.52

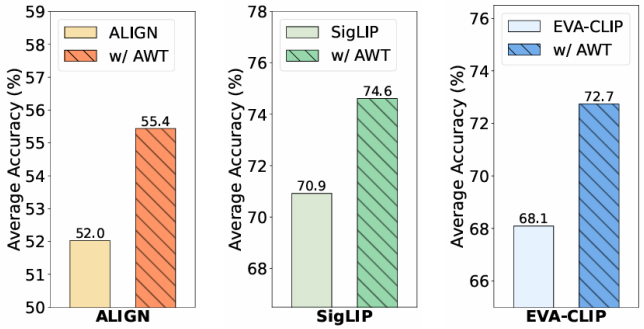
(c) Number of generated **class descriptions** M.

M	OOD	Avg.(14)
2	61.58	66.67
5	62.87	67.55
10	63.42	68.53
25	64.07	68.98
50	64.43	69.59
100	**64.46**	**69.73**

  - Applicable to any dual encoder VLM



(a) AWT across different backbones.



(b) AWT generalizes to different VLMs.

AWT[1)]

- Conclusion
  - Contributions
    - Backbone model에 상관없이 적용할 수 있는 training-free method 제시
    - 모델의 해석 및 활용 용이성
    - Class descriptions를 모두 활용하는 수 있는 transport 방식 제안
  - Limitations
    - View 개수 증가에 따라 inference시 높은 computational cost을 요함
    - 단순 Random Resize Crop으로 인해 중복되는 view가 생성됨
    - Training을 통한 성능 개선 방법은 제시되지 않음

감사합니다