특집논문 (Special Paper)
방송공학회논문지 제30권 제4호, 2025년 7월 (JBE Vol.30, No.4, July 2025)
https://doi.org/10.5909/JBE.2025.30.4.571
ISSN 2287-9137 (Online) ISSN 1226-7953 (Print)

메타학습을 이용한 시각-언어 모델 프롬프트 튜닝

김 도 현[a], 백 성 용[a]‡

Prompt Tuning for Vision-Language Models via Meta-Training

Dohyeon Kim[a] and Sungyong Baik[a]‡

요 약

최근, CLIP과 같은 대규모 사전학습된 시각-언어 모델(Vision-Language Model)을 다양한 다운스트림 태스크에 적용한 연구들이 우수한 성능을 보이고 있다. 특히 소수의 샘플만을 활용하는 이미지 분류(Low-shot Image Classification)에서는, 연속적인 프롬프트 벡터를 최적화하는 방식이 주목받고 있으나, 일반화 성능이 낮다는 한계가 있다. 최근 연구들은 해당 문제를 해결하기 위해서 추가적인 모델 구조나 알고리즘을 도입하지만, 이는 효율성이 저하되는 단점을 지닌다. 본 논문에서는 일반화 성능을 효율적으로 높이기 위해, 멀티모달(Multi-modal) 표현을 활용한 메타러닝(Meta-Learing) 알고리즘으로 프롬프트 벡터를 최적화시키는 방법을 제안한다. 제안하는 방법은 학습된 도메인과 신규 도메인 모두 고려한 성능에서 다른 모델들에 비해 약 9.6% 이상의 정확도 향상을 보이며, 추가적인 메모리나 지연시간 없이 제로샷 추론이 가능하다.

Abstract

Recently, applying large pre-trained vision-language models such as CLIP to various downstream tasks has shown good performance. In low-shot image classification, simply optimizing continuous prompt vectors emerged, but has the limitation of low generalizability. Recent research introduces additional structures and algorithms to solve this problem, but there is a disadvantage of inefficiency. So, we propose a meta-training framework utilizing multi-modal features to optimize the prompt vectors. Our proposed method achieves over a 9.6% accuracy improvement compared to other models when evaluated on both seen and unseen domains comprehensively, which has no any additional memory overhead or inference latency for zero-shot inference.

Keyword : Vision-Language Models, Prompt Tuning, Meta-Learning, Low-shot Image Classification

a) 한양대학교 데이터사이언스학부(Department of Data Science, Hanyang University)
‡ Corresponding Author : 백성용(Sungyong Baik)
E-mail: dsybaik@hanyang.ac.kr
Tel: +82-2-2220-2524
ORCID: https://orcid.org/0000-0001-5702-4618

※ This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2020-0-01373, Artificial Intelligence Graduate School Program(Hanyang University))
· Manuscript May 22, 2025; Revised July 2, 2025; Accepted July 7, 2025.

Ⅰ. 서 론

CLIP[1], ALIGN[2]과 같은 대규모 사전학습 시각-언어 모델(Vision-Language Model)의 등장으로, 텍스트-이미지 쌍의 표현을 다양한 다운스트림 태스크에 활용하는 것이 가능해졌다. 예를 들어, CLIP[1]의 경우, 각 레이블을 포함한 프롬프트를 작성한 후, 텍스트 인코더를 통해 생성된 텍스트 표현과 이미지 인코더를 통해 생성된 이미지 표현 간의 유사도를 계산함으로써 분류 태스크를 수행할 수 있다. 그러나 사용된 프롬프트에 따라 성능이 크게 달라지며, 해당 태스크에 적합한 프롬프트를 탐색하는 과정은 중요한 요소로 작용한다.

우선, 하드 프롬프트(Hard Prompt)라고 불리는 수작업 기반의 프롬프트는 여러 가지 한계를 지닌다. 하드 프롬프트가 특정 다운스트림 태스크(Downstream Task)에 최적화되어 있는지 확신하기 어렵고, 이산적인 값으로 구성되어 있어 표현력 측면에서도 제약이 존재한다. 이에 따라, 연속적인 실숫값으로 구성된 소프트 프롬프트(Soft Prompt)를 구성하고, 소량의 데이터를 이용하여 학습시킴으로써 각 다운스트림 태스크에 최적화된 프롬프트를 얻고자 하는 연구가 진행되었으며, 이를 기반으로 CoOp[3]이 제안되었다. CoOp[3]은 학습된 도메인 내에서는 기존 CLIP[1]의 제로샷(Zero-shot) 이미지 분류 성능을 크게 상회하는 결과를 보였으나, 유사한 다른 도메인에서는 오히려 CLIP[1]보다 낮은 성능을 보여주었다. 이는 소량의 학습 데이터로 소프트 프롬프트를 학습하는 과정에서 과적합(Overfitting)이 발생하여, 대규모 사전학습된 시각-언어 모델이 보유한 일반화된 지식이 손실되었기 때문이며, 이러한 문제를 해결하고자 다양한 후속 연구들이 수행되고 있다.

CoCoOp[4]는 기존 CoOp[3]의 소프트 프롬프트에 경량 메타 네트워크(Meta-Network)를 통과한 이미지 표현을 추가하여 입력 조건에 따른 프롬프트를 생성하는 방법을 제안하였다. Argue[5]와 LLaMP[6]는 대형 언어 모델(Large Language Model)을 활용하여 각 클래스명에 대응하는 정교한 시각적 특징을 사용하는 방식을 제안하였다. ProMetaR[7]는 최적화 기반의 메타러닝 알고리즘을 이용하여 정규화 항과 프롬프트를 동시에 업데이트하는 방법을 제안하였고, PromptKD[8]는 레이블이 없는 데이터와 프롬프트를 이용하여 교사 모델(Teacher Model)로부터 생성된 텍스트 표현과 이미지 표현 간의 확률 분포를 학습함으로써 학생 모델(Student Model)이 이를 모방할 수 있도록 하는 방법을 제안하였다.

이와 같은 기존 접근 방식들은 프롬프트를 효과적으로 학습하면서 일반화 성능을 유지하였으나, 추론 시에 추가적인 모델이나 알고리즘이 적용되어 효율성이 저하된다는 한계를 지니고 있다. 이러한 문제를 해결하기 위해, 본 논문에서는 추론 단계에서도 CLIP[1]의 제로샷 이미지 분류처럼 간단히 적용하면서도 일반화 성능을 높인 메타학습 프레임워크(Meta-Training Framework)를 제안한다. 본 프레임워크는 텍스트와 이미지의 멀티모달(Multi-modal) 표현과 함께 메타러닝(Meta-Learning) 알고리즘을 적용한다.

Ⅱ. 관련 연구

1. 시각-언어 모델 (Vision-Language Models)

시각-언어 모델은 이미지와 텍스트 임베딩을 동일한 임베딩 공간에 정렬(Alignment)하는 것을 목표로 한다. 최근의 시각-언어 모델들은 이미지 인코더와 텍스트 인코더를 동시에 학습하여 두 모달리티(Modality) 간의 간극을 연결하도록 설계되었다. 이러한 시각-언어 모델은 제로샷 이미지 분류 태스크에서 우수한 성능을 보였으며, 이러한 성과에는 Transformer[9] 기반의 인코더(Encoder) 구조, 대조 학습(Contrastive Learning) 기법[10-12], 그리고 대규모 이미지-텍스트 데이터셋이 중요한 역할을 하였다. 대표적인 모델로는 CLIP[1]과 ALIGN[2]이 있으며, CLIP[1]은 약 4억 개의 이미지-텍스트 쌍을 활용하여 이미지-텍스트 쌍 간 임베딩의 대조 손실(Contrastive Loss)을 기반으로 학습되었다. ALIGN[2]은 정제된 10억 개의 이미지-텍스트 쌍을 사용하였고, 이미지-텍스트 및 텍스트-이미지 분류 손실을 포함한 대조 손실을 이용하여 학습되었다.

2. 프롬프트 튜닝 (Prompt Tuning)

프롬프트는 BERT[13], GPT[14-16]와 같은 사전학습 언어

모델(Language Model)이 특정 다운스트림 태스크에서 우수한 성능을 발휘하도록 유도하는 입력 텍스트를 의미한다. 다운스트림 태스크의 성능은 프롬프트에 따라 크게 달라지며, 프롬프트 튜닝은 사전학습된 언어 모델과 다운스트림 태스크에 적합한 프롬프트를 어떻게 설계하고 최적화시킬 것인가에 대한 연구 분야이다. Jiang[17] 등은 프롬프트를 생성한 뒤 마이닝(Mining) 또는 파라프레이징(Paraphrasing)을 통해 최적의 프롬프트를 선택하는 프레임워크를 제안하였으며, AutoPrompt[18]는 그래디언트(Gradient) 기반의 방법으로 그래디언트 변화가 큰 토큰을 선택하여 프롬프트를 구성하는 방식을 제안하였다. 소프트 프롬프트 기반 방식들[19-22]에서는 프롬프트를 연속적인 값을 가진 벡터들의 집합으로 정의하고, 기존의 손실 함수(Loss Function)를 통해 직접 학습하는 방식이 적용된다. 이러한 프롬프트 튜닝 방식은 최근 컴퓨터 비전 분야로도 확장되고 있으며, CoOp[3], CoCoOp[4] 등이 대표적이다. 이 외에도 대형 언어 모델을 활용한 Argue[5], LLaMP[6], 메타러닝 알고리즘을 적용한 ProMetaR[7], 지식 증류(Knowledge Distillation)를 활용한 PromptKD[8] 등의 연구가 진행되고 있다.

3. 메타러닝 (Meta-Learning)

사람은 지금까지 학습한 지식을 바탕으로 새로운 상황에 빠르게 적응하는 능력을 지닌다. 메타러닝은 이러한 인간의 학습 방식에 착안하여, 소량의 데이터만으로도 새로운 태스크에 빠르게 적응할 수 있도록 하는 학습 알고리즘이다. 메타러닝은 크게 최적화 기반(Optimization-based) 접근 방식과 '다른 하나는' 삭제 거리 기반(Metric-based) 접근 방식으로 나눌 수 있다. 최적화 기반 접근 방식은 특정 태스크를 수행하는 학습자(Learner)의 가중치를 학습시키는 메타 학습자(Meta-Learner)를 학습하는 것을 목표로 한다. 대표적인 예로는 Meta-SGD[23], Meta-LSTM[24] 등이 있으며, 메타 초기 가중치(Meta-Initialization)를 학습하는 MAML[25] 또한 속한다. 한편, 거리 기반 접근 방식은 태스크 공간 내에서 데이터 간 관계를 나타내는 함수를 학습하는 것을 목표로 한다. 이는 각 데이터를 특정 함수를 통해 임베딩 공간으

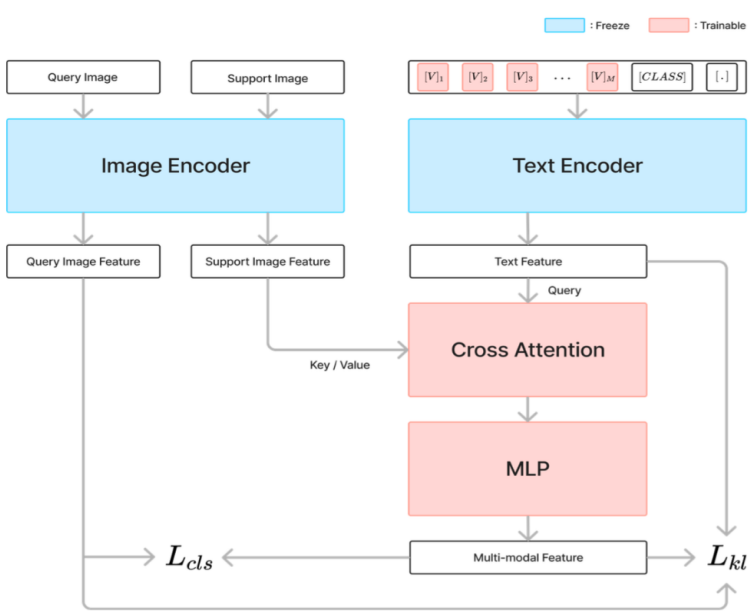

그림 1. 전체적인 메타학습 프레임워크 구조도 $[V]$: 학습 가능한 소프트 프롬프트 벡터, $[CLASS]$: 클래스 이름 토큰 벡터
Fig. 1. Overall our meta-training framework $[V]$: Trainable soft prompt vector, $[CLASS]$: Class name token vector

로 매핑한 뒤, 해당 공간 내에서 유클리드 거리(Euclidean Distance)나 코사인 유사도(Cosine Similarity)와 같은 메트릭(Metric)을 기준으로 같은 클래스 간 거리는 작아지도록, 다른 클래스 간 거리는 커지도록 학습시킨다. 대표적인 모델로는 Relation Network(RN)[26], ProtoNet[27] 등이 있다.

Ⅲ. 제안 방법

본 프레임워크는 그림 1에 제시되어 있으며, 이를 이해하기 위한 사전 지식은 3.1절에서 설명한다. 이어서, 제안하는 모델의 구조는 3.2절에서 구체적으로 기술하며, 3.3절에서는 본 모델이 어떠한 손실 함수를 통해 학습되는지를 설명한다. 마지막 3.4절에서는 학습된 모델이 어떻게 제로샷 추론(Zero-shot Inference)을 수행하는지를 다룬다.

1. 사전 지식 (Preliminaries)

CLIP[1]에서의 대조 학습[10]: CLIP[1]에서는 이미지 인코더와 텍스트 인코더를 학습하기 위해, 이미지-텍스트 쌍 간의 대조 손실을 활용한다. 손실 함수는 양의 이미지-텍스트 쌍 간의 코사인 유사도는 높이고, 음의 쌍 간의 유사도는 낮추도록 설정된다. 클래스의 개수를 N, 이미지 인코더로부터 얻은 이미지 표현을 $\boldsymbol{x}$, 텍스트 인코더로부터 얻은 N개의 텍스트 표현 집합을 $\{\boldsymbol{w}_i\}_{i=1}^{N}$라고 할 때, 각 클래스에 대한 예측 확률은 다음과 같이 정의된다:

$$p(y|\boldsymbol{x}) = \frac{\exp(sim(\boldsymbol{x}, \boldsymbol{w}_y)/\tau)}{\sum_{i=1}^{N} \exp(sim(\boldsymbol{x}, \boldsymbol{w}_i)/\tau)} \quad (1)$$

여기서 sim(· , ·)은 코사인 유사도를 의미하며, τ는 하이퍼파라미터이다. 모델은 예측 확률 분포를 이용하여 계산된 크로스 엔트로피 손실(Cross Entropy Loss)을 통해 업데이트되며, 이미지 축과 텍스트 축 모두에서 손실이 계산된다.

Prototypical Network[27]: ProtoNet이라고 부르며, 메타러닝에서 대표적인 거리 기반 접근 방식 중 하나이다. 메타러닝에서는 데이터셋이 여러 개의 태스크로 구성되며, 각 태스크는 Support 데이터와 Query 데이터로 구성된다. X_i^s, X_i^q가 각각 클래스에 대한 Support 데이터, Query 데이터라고 가정하면 각 태스크는 $T = \{X_i^s, X_i^q\}_{i=1}^{N}$로 표현되고, 이미지 인코더 f_θ를 통해 Support 데이터 표현들을 구한 후 각 클래스마다 평균을 내면 각 클래스의 프로토타입(Prototype)을 계산할 수 있다:

$$c_i = \frac{1}{N} \sum_{x_i^s \in X_i^s} f_\theta(x_i^s) \quad (2)$$

이후, 클래스의 프로토타입을 이용하여 다음과 같이 데

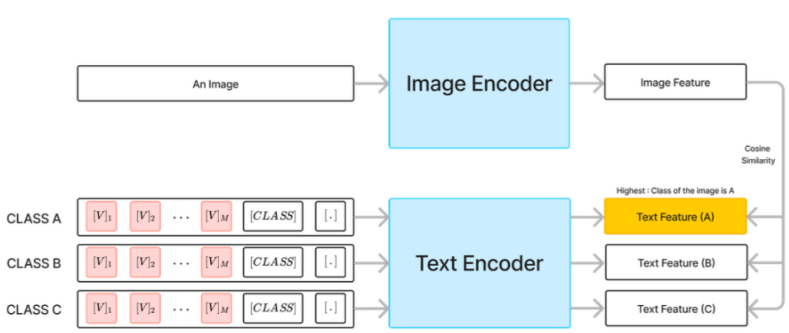

그림 2. 제로샷 이미지 분류 절차 구조도 $[V]$: 학습된 소프트 프롬프트 벡터, $[CLASS]$: 클래스 이름 토큰 벡터
Fig. 2. The procedure of zero-shot image classification $[V]$: Trained soft prompt vector, $[CLASS]$: Class name token vector

이터의 각 클래스에 대한 예측 확률을 정의할 수 있다:

$$p(y|\boldsymbol{x}) = \frac{\exp(-dist(\boldsymbol{x}, \boldsymbol{c}_y))}{\sum_{i=1}^{K} \exp(-dist(\boldsymbol{x}, \boldsymbol{c}_i))} \quad (3)$$

여기서 dist($\cdot, \cdot$)는 유클리드 거리(Euclidean Distance)를 의미하며, 모델은 Query 데이터의 예측 확률 분포로 계산된 크로스 엔트로피 손실을 통해 업데이트된다.

2. 모델 구조 (Architecture)

제안하는 모델 구조는 비교적 단순하다. CoOp[3]과 마찬가지로 학습 가능한 소프트 프롬프트 벡터, 이미지 인코더, 텍스트 인코더로 구성되며, 여기에 추가적으로 크로스 어텐션(Cross-Attention)[9] 모듈과 다층 퍼셉트론(MLP) 모듈이 존재한다. 이미지 인코더와 텍스트 인코더는 사전학습된 CLIP[1]을 그대로 사용하며, 학습 과정에서는 해당 가중치를 고정시킨다. 소프트 프롬프트 벡터의 차원은 텍스트 인코더의 입력 차원과 동일하게 설정되며, 벡터의 개수는 하이퍼파라미터로 자유롭게 조정 가능하다. 크로스 어텐션[9] 모듈에서는 이미지 인코더로부터 추출된 표현을 Key와 Value로 사용하고, 텍스트 인코더로부터 추출된 표현을 Query로 사용하며, 잔차 연결(Residual Connection)도 함께 적용된다. 이후 다층 퍼셉트론 모듈을 지나 최종적으로 멀티모달 표현을 생성한다. 각 모듈의 세부적인 역할은 다음 절인 3.3절에서 자세히 설명한다.

3. 메타학습 (Meta-Training)

본 연구에서는 소프트 프롬프트 벡터, 크로스 어텐션[9] 모듈, 그리고 다층 퍼셉트론 모듈을 ProtoNet[27]과 같은 방식으로 메타학습을 진행한다. 클래스 프로토타입 생성을 위해, 텍스트 인코더로부터 얻은 텍스트 표현과 이미지 인코더로부터 얻은 Support 데이터 표현을 크로스 어텐션[9] 모듈과 다층 퍼셉트론 모듈을 통해 결합하는 방식을 사용한다. 각 Query 이미지 샘플에 대한 예측 확률은 식 (1)을 따르며, 해당 확률 분포를 기반으로 다음과 같이 크로스 엔트로피 손실을 계산한다:

$$L_{cls} = -\frac{1}{N} \sum_{i=1}^{N} \log p(\boldsymbol{x}^i | \boldsymbol{c}_y^i) \quad (4)$$

이러한 메타학습 방식은 소수의 샘플만으로도 클래스의 일반화된 표현을 효과적으로 생성할 수 있도록 도와주어, 학습된 도메인뿐만 아니라 보지 못한 새로운 도메인에서도 퓨샷 데이터만으로 우수한 성능을 보이게 한다. 그러나 추론을 수행할 때 클래스 프로토타입을 생성하기 위해 레이블링된 데이터 샘플이 반드시 필요하다는 한계점이 존재하는데, 이는 기존의 CLIP[1]이나 CoOp[3]과 같이 추론 시 레이블링된 데이터 없이도 작동 가능한 제로샷 추론이 불가능하다는 점에서 중요한 제약 사항으로 작용한다. 이에 대한 해결 방안으로 본 연구에서는 추론 단계에서 멀티모달 표현이 아닌 텍스트 표현만을 사용하는 방식을 채택한다. 하지만, 본 모델은 학습 시 크로스 엔트로피 손실을 멀티모달 표현을 기반으로 계산하므로, 텍스트 표현만을 사용하여 제로샷 추론을 진행할 경우 성능이 저하될 가능성이 있다. 이를 보완하기 위해, 멀티모달 표현과 텍스트 표현 각각을 사용하여 얻은 예측 확률 분포가 유사하도록 모델을 학습시키며, 이때 사용하는 KL Divergence 손실은 다음과 같다:

$$L_{kl} = KL_{divergence}(P_{mm} | P_{text}) \quad (5)$$

여기서 P_{mm}은 멀티모달 표현을 사용한 예측 확률 분포, P_{text}는 텍스트 표현을 사용한 예측 확률 분포를 의미하며, 이는 멀티모달 표현과 텍스트 표현이 유사하도록 만드는 역할을 한다. 따라서 최종 손실 함수는 다음과 같이 정의된다:

$$L = L_{cls} + \lambda L_{kl} \quad (6)$$

4. 제로샷 추론 (Zero-shot Inference)

제로샷 추론 절차는 그림 2에 제시되어 있다. 메타학습 때 사용된 크로스 어텐션[9] 모듈, 그리고 다층 퍼셉트론 모듈은 더 이상 사용하지 않으며, 이미지 인코더를 통해 추출된 이미지 표현과 학습된 소프트 프롬프트 벡터, 텍스트 인코더를 통해 추출된 텍스트 표현만을 사용하여 추론을 진

행한다. 이미지 표현과의 코사인 유사도가 가장 높은 텍스트 표현을 가지는 레이블을 선택함으로써 최종적으로 예측을 수행할 수 있다.

Ⅳ. 실험 및 결과

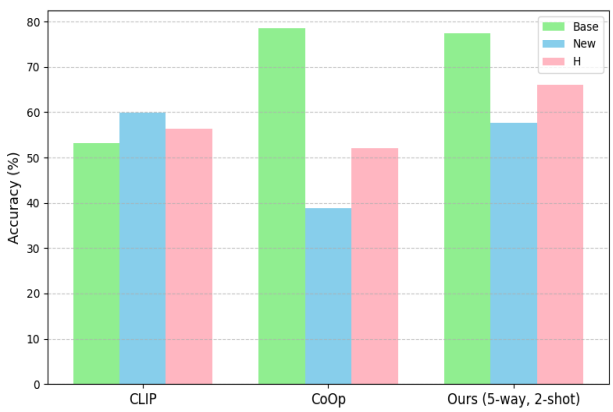

그림 3. '다른' 삭제 베이스라인 모델들(CLIP[1], CoOp[3])과의 성능 비교. Base: Base 클래스 데이터셋(테스트)에서의 정확도, New: New 클래스 데이터셋(테스트)에서의 정확도, H: Base와 New 성능의 조화 평균
Fig. 3. Performance comparison to 'other' 삭제 baseline models(CLIP[1], CoOp[3]). Base: Accuracy on Base class dataset(test), New: Accuracy on New class dataset(test), H: Harmonic mean of Base and New perfomance

표 1. 메타학습 설정에서의 N-way에 따른 본 모델 성능 비교. Base: Base 클래스 데이터셋(테스트)에서의 정확도, New: New 클래스 데이터셋(테스트)에서의 정확도, H: Base와 New 성능의 조화 평균
Table 1. Performace comparison between various N-ways in the meta-training setting. Base: Accuracy on Base class dataset(test), New: Accuracy on New class dataset(test), H: Harmonic mean of Base and New perfomance

	Base	New	H
Ours (5-way)	77.31%	57.65%	66.02%
Ours (10-way)	76.39%	57.27%	65.48%
Ours (20-way)	76.62%	57.27%	65.33%

1. 베이스라인 (Baselines)

본 모델과의 성능 비교를 위해 CLIP[1]과 CoOp[3]을 베이스라인으로 활용하였다. 두 모델 모두 이미지 인코더 아키텍처로 ViT[28], 특히 ViT-B/16을 사용하는 사전학습된 CLIP 인코더를 사용한다. CLIP은 추가 학습을 진행하지 않으며, CoOp[3]은 "a photo of a"라는 텍스트의 임베딩 표현으로 초기화된 4개의 소프트 프롬프트 벡터만을 200 에폭(Epoch) 동안 학습한다. 이때 배치(Batch) 크기는 32, 옵티마이저(Optimizer)는 SGD, 모멘텀(Momentum)은 0.9, 초기 학습률(Learing Rate)은 2e-3, 가중치 감쇠(Weight Decay)는 5e-4이며, 코사인 학습률 스케줄러(Cosine Learning Rate Scheduler)를 사용한다. 텍스트 표현을 생성할 때, CLIP[1]은 프롬프트 "a photo of a [class]"를 사용하지만, CoOp[3]은 소프트 프롬프트 벡터와 클래스 이름을 함께 사용한다.

2. 데이터셋 (Dataset)

모델 성능 평가를 위해 이미지 인식 데이터셋을 사용하였으며, 구체적으로 텍스처 분류를 위한 DTD[29] 데이터셋을 활용하였다. 해당 데이터셋은 총 47개의 클래스로 구성되어 있으며, 본 연구에서는 학습된 도메인을 평가하기 위한 Base 클래스(24개)와 신규 도메인을 평가하기 위한 New 클래스(23개)로 분리한다. 베이스라인 모델들과 본 모델은 Base 클래스 각각에 대해 무작위로 샘플링된 16개의 데이터로 학습되며, 데이터 증강(Data Augmentation) 기법으로는 Random Resized Crop과 Random Horizontal Flip을 적용하였다. 메타 데이터셋(Meta-Dataset)에서는 N-way K-shot 설정을 기반으로 태스크를 구성하며, 각 태스크마다 N개의 클래스를 무작위로 샘플링한 후, 클래스당 16개의 데이터를 샘플링하여 그 중 K개를 Support 데이터로, 나머지를 Query 데이터로 사용한다.

3. 학습 세부사항 (Training Details)

본 모델의 하이퍼파라미터(Hyperparameter)는 CoOp[3]과 최대한 동일하게 설정하였다. CoOp[3]과 마찬가지로, 이미지 인코더로 VIT-B/16을 사용하는 사전학습된 CLIP[1] 인코더를 사용하며, 소프트 프롬프트 벡터 또한 같은 방식으로 초기화한다. 에폭 수는 100으로 설정하였고, 나머지

하이퍼파라미터는 CoOp[3]과 동일하다. 메타학습과 관련된 하이퍼파라미터로는, 에폭당 반복(Iteration) 수를 10으로 설정하였으며, 메타 배치 크기(Meta Batch Size)는 4, Support 데이터 샘플 수는 2로 설정하였다. 선택되는 클래스 수는 각각 5, 10, 20으로 설정하여 다양한 실험을 진행하였다. 마지막으로 손실 함수에 있어 식 (6)의 λ 값은 0.2로 설정하였다.

4. 결과 분석 (Analysis)

그림 3은 베이스라인 모델들과 본 모델 간의 성능 비교 결과를 나타낸다. 본 모델은 Base 클래스에서 CoOp[3]보다, New 클래스에서는 CLIP[1]보다 다소 낮은 성능을 보였으나, 두 설정에서의 조화 평균(Harmonic Mean)은 다른 모델들에 비해 현저히 높게 나타났다. 이는 본 모델이 일반화된 지식을 유지하면서도 신규 도메인에 효과적으로 적응하고 있음을 의미한다. 또한, 다양한 N-way 설정 하에서 동일한 K-shot(2-shot)으로 실험을 진행하였으며, 그 결과는 표 1에 정리하였다. 큰 차이는 없었으나, N-way가 증가함에 따라 성능이 소폭 감소하는 경향을 보였다. 이는 N-way가 증가할수록 샘플링되는 태스크가 24개 클래스를 분류하는 원래 태스크에 더 가까워지며, 이에 따라 태스크 간 분산이 줄어들어 일반화 성능이 다소 감소한 것으로 판단된다.

5. KL Divergence 손실의 유효성 (Validity of KL Divergence Loss)

그림 4를 통해 본 모델의 학습에 사용된 KL Divergence 손실의 효과를 확인할 수 있다. 멀티모달 표현을 사용하여 예측을 수행한 경우와 텍스트 표현을 사용하여 예측을 수행한 경우를 비교하였을 때, 학습 초기에는 정확도 차이가 약 6% 이상이었지만, 학습이 더 진행됨에 따라 정확도 차이가 약 1~2% 정도로 줄어들었다. 이는 추론 때 멀티모달 표현을 만들기 위해 사용되는 추가적인 Support 데이터 없이 텍스트 표현만으로도 비슷한 성능을 낼 수 있음을 의미하며, 다른 베이스라인 모델들과 같이 제로샷 추론이 가능함을 보여준다.

V. 결 론

본 논문에서는 멀티모달 표현을 활용한 메타러닝 알고리즘으로 프롬프트 벡터를 최적화시키는 방법을 제안한다. 제안된 방법은 기존의 방법들보다 도메인에 대한 적응 및 일반화 능력 측면에서 우수한 성능을 나타낸다. 특히, 기존 메타러닝 기반 방법들이 추론 단계에서 레이블링된 소량의 데이터를 요구하는 반면, 본 모델은 추론 단계에서 해당 제

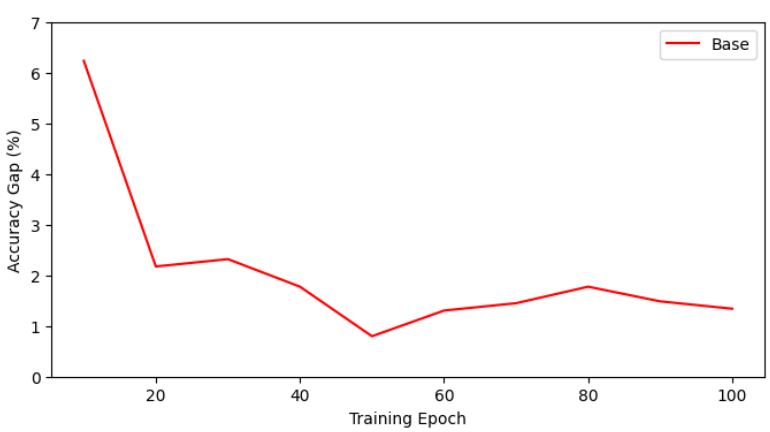

그림 4. Base 클래스 데이터셋(테스트)에서 멀티모달 표현과 텍스트 표현 각각 예측을 수행하였을 경우, 학습 에폭에 따른 정확도 차이
Fig. 4. Epoch-wise differences in acuuracy between multi-modal and text features on Base class dataset(test)

약 없이 사용 가능하며, 추가적인 모델이나 알고리즘을 필요로 하지 않아 다른 방법들에 비해 높은 효율성을 지닌다.

참 고 문 헌 (References)

[1] A. Radford et al., "Learning Transferable Visual Models from Natural Language Supervision," *Proceedings of the 38th International Conference on Machine Learning (ICML)*, Virtual, pp. 8748 – 8763, 2021.
doi: https://doi.org/10.48550/arXiv.2103.00020

[2] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. V. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision," *Proceedings of the 38th International Conference on Machine Learning (ICML)*, Virtual, pp. 4904 – 4916, 2021.
doi: https://doi.org/10.48550/arXiv.2102.05918

[3] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to Prompt for Vision Language Models," *International Journal of Computer Vision*, Vol. 130, No. 9, pp. 2337 – 2348, Sept. 2022.
doi: https://doi.org/10.1007/s11263-022-01653-1

[4] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional Prompt Learning for Vision Language Models," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, LA, USA, pp. 16816 – 16825, 2022.
doi: https://doi.org/10.1109/CVPR52688.2022.01631

[5] X. Tian, S. Zou, Z. Yang, and J. Zhang, "ArGue: Attribute Guided Prompt Tuning for Vision Language Models," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, pp. 28578 – 28587, 2024.
doi: https://doi.org/10.1109/CVPR52733.2024.02700

[6] Z. Zheng, J. Wei, X. Hu, H. Zhu, and R. Nevatia, "Large Language Models Are Good Prompt Learners for Low Shot Image Classification," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, pp. 28453 – 28462, 2024.
doi: https://doi.org/10.1109/CVPR52733.2024.02688

[7] J. Park, J. Ko, and H. J. Kim, "Prompt Learning via Meta-Regularization," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, pp. 26930 – 26940, 2024.
doi: https://doi.org/10.1109/CVPR52733.2024.02544

[8] Z. Li, X. Li, X. Fu, X. Zhang, W. Wang, S. Chen, and J. Yang, "PromptKD: Unsupervised Prompt Distillation for Vision Language Models," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, pp. 26617 – 26626, 2024.
doi: https://doi.org/10.1109/CVPR52733.2024.02513

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," *Advances in Neural Information Processing Systems 30*, Long Beach, CA, USA, pp. 5998 – 6008, 2017.
doi: https://doi.org/10.48550/arXiv.1706.03762

[10] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," *Proceedings of the 37th International Conference on Machine Learning (ICML)*, Virtual, pp. 1597 – 1607, 2020.
doi: https://doi.org/10.48550/arXiv.2002.05709

[11] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum Contrast for Unsupervised Visual Representation Learning," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Virtual, pp. 9729 – 9738, 2020.
doi: https://doi.org/10.1109/CVPR42600.2020.00975

[12] O. Hénaff, A. Srinivas, J. De Fauw, A. Razavi, C. Doersch, S. M. A. Eslami, and A. van den Oord, "Data Efficient Image Recognition with Contrastive Predictive Coding," *Proceedings of the 37th International Conference on Machine Learning (ICML)*, Virtual, pp. 4182 – 4192, 2020.
doi: https://doi.org/10.48550/arXiv.1905.09272

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*, Minneapolis, MN, USA, vol. 1 (Long and Short Papers), pp. 4171 – 4186, 2019.
doi: https://doi.org/10.18653/v1/N19-1423

[14] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving Language Understanding by Generative Pre-Training," OpenAI, Jun. 11, 2018.

[15] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language Models Are Unsupervised Multitask Learners," OpenAI Blog, Feb. 14, 2019.

[16] T. B. Brown et al., "Language Models Are Few-Shot Learners," *Advances in Neural Information Processing Systems 33*, Virtual, pp. 1877 – 1901, 2020.
doi: https://doi.org/10.48550/arXiv.2002.05709

[17] Z. Jiang, F. F. Xu, J. Araki, and G. Neubig, "How Can We Know What Language Models Know?," *Transactions of the Association for Computational Linguistics*, Vol. 8, pp. 423 – 438, 2020.
doi: https://doi.org/10.1162/tacl_a_00324

[18] T. Shin, Y. Razeghi, R. L. Logan IV, E. Wallace, and S. Singh, "AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts," *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (EMNLP), pp. 4222 – 4235, 2020.
doi: https://doi.org/10.18653/v1/2020.emnlp-main.346

[19] X. L. Li and P. Liang, "Prefix Tuning: Optimizing Continuous Prompts for Generation," *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL IJCNLP 2021)*, Virtual, pp. 4582 – 4597, 2021.
doi: https://doi.org/10.18653/v1/2021.acl-long.353

[20] X. Liu, Y. Zheng, Z. Du, M. Ding, Y. Qian, Z. Yang, and J. Tang, "GPT

Understands, Too," *AI Open*, Vol. 5, pp. 208–215, Aug. 2023 (published 2024).
doi: https://doi.org/10.1016/j.aiopen.2023.08.012

[21] X. Liu, K. Ji, Y. Fu, W. L. Tam, Z. Du, Z. Yang, and J. Tang, "P Tuning: Prompt Tuning Can Be Comparable to Fine tuning Universally Across Scales and Tasks," *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, Dublin, Ireland, pp. 61–68, 2022.
doi: https://doi.org/10.18653/v1/2022.acl-short.8

[22] B. Lester, R. Al Rfou, and N. Constant, "The Power of Scale for Parameter Efficient Prompt Tuning," *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online and Punta Cana, Dominican Republic, pp. 3045–3059, 2021.
doi: https://doi.org/10.18653/v1/2021.emnlp-main.243

[23] Z. Li, F. Zhou, F. Chen, and H. Li, "Meta-SGD: Learning to Learn Quickly for Few-Shot Learning," *arXiv preprint arXiv:1707.09835*, Jul. 31, 2017.
doi: https://doi.org/10.48550/arXiv.1707.09835

[24] S. Ravi and H. Larochelle, "Optimization as a Model for Few-Shot Learning," *International Conference on Learning Representations (ICLR)*, Toulon, France, 2017. Available: https://openreview.net/forum?id=rJY0-Kcll

[25] C. Finn, P. Abbeel, and S. Levine, "Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks," *Proceedings of the 34th International Conference on Machine Learning (ICML)*, Sydney, Australia, pp. 1126–1135, 2017.
doi: https://dl.acm.org/doi/10.5555/3305381.3305498

[26] S. Sung, Y. Flood, and et al., "Learning to Compare: Relation Network for Few-Shot Learning," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, pp. 1199–1208, 2018.
doi: https://doi.org/10.1109/CVPR.2018.00131

[27] J. Snell, K. Swersky, and R. Zemel, "Prototypical Networks for Few-shot Learning," *Advances in Neural Information Processing Systems 30*, Long Beach, CA, USA, pp. 4077–4087, 2017.
doi: https://dl.acm.org/doi/10.5555/3294996.3295163

[28] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale," *International Conference on Learning Representations (ICLR)*, Virtual Event, Austria, 2021.
doi: https://doi.org/10.48550/arXiv.2010.11929

[29] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi, "Describing Textures in the Wild," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Columbus, Ohio, USA, pp. 3606–3613, 2014.
doi: https://doi.org/10.1109/CVPR.2014.461

저 자 소 개

김 도 현

- 2020년 ~ 현재 : 한양대학교 데이터사이언스학부 학사과정
- ORCID : https://orcid.org/0009-0006-2935-607X
- 주관심분야 : 메타러닝, 퓨샷러닝, 컴퓨터비전

백 성 용

- 2022년 3월 : 서울대학교 박사
- 2022년 3월 ~ 현재 : 한양대학교 데이터사이언스학부 조교수
- ORCID : https://orcid.org/0000-0001-5702-4618
- 주관심분야 : 범용인공지능, 멀티모달 인공지능, 컴퓨터비전, 퓨샷러닝, 메타러닝