단일 양성 레이블 데이터 환경에서의 학습 비의존 Few-shot 적응 기반 다중 레이블 분류 모델

심훈보, 곽민준, 정규원, 전창재*
세종대학교

hunbo00@sju.ac.kr, gminjun76@gmail.com, kwsoosoo0815@gmail.com, *cchun@sejong.ac.kr

Multi-label Classification Model based on Training-free Few-shot Adaptation for Single Positive Label Dataset

Hunbo Shim, Minjun Kwak, Kyuwon Jung, Chang-Jae Chun*
Sejong Univ.

요 약

최근 대규모 비전-언어 모델인 CLIP은 우수한 zero-shot 성능을 보이나, 한 이미지 내 다수의 객체가 존재하더라도 텍스트는 주로 단일 객체만을 지칭하는 데이터로 학습된 특성상, 다중 객체가 혼재된 환경에서는 주변부 객체나 작은 객체를 식별하는 데 한계가 있다. 이를 해결하기 위한 전체 데이터 레이블링은 막대한 구축 비용이 소요된다. 이에 본 논문에서는 비용 효율적인 단일 양성 레이블(Single Positive Label, SPL) 제약 조건 하에서, 학습 비의존 few-shot 적응 기법인 Tip-Adapter를 활용하여 다중 레이블 분류 성능을 확인했다. 본 연구는 클래스 별 독립적인 캐시 구조를 제안하여 전역 특징을 효율적으로 활용했다. 실험 결과, SPL 기반의 few-shot 캐시와 선형 결합만으로도 추가 학습 없이 기존 zero-shot 대비 유의미하게 향상된 다중 객체 추론 성능을 입증했다.

Ⅰ. 서론

대규모 이미지-텍스트 쌍으로 사전 학습된 CLIP (Contrastive Language-Image Pre-training)[1]의 등장으로 별도의 학습 없는 zero-shot 추론이 가능해졌다. 그러나 CLIP은 비록 한 이미지 속에 여러 클래스가 혼재해 있더라도 텍스트는 주로 가장 두드러진 하나의 메인 객체만을 묘사하는 데이터 쌍을 통해 학습되었다. 이로 인해 모델은 이미지 내 가장 지배적인 객체의 특징에만 편향적으로 학습되는 결과를 초래했으며, 결과적으로 크기가 작거나 주변부에 위치한 객체들을 포착하는 데 한계를 보인다. 이를 극복하기 위해 모델을 다중 레이블 데이터셋에 맞춰 미세 조정하는 방법이 있으나, 이를 위해서는 이미지 내 존재하는 모든 객체를 레이블링하는 과정이 선행되어야 한다. 하지만 이는 데이터 구축에 막대한 비용과 시간이 소요되므로 현실적으로 적용하기 어렵다. 이에 본 논문에서는 현실적으로 확보가 용이하고 보편적인 단일 양성 레이블 데이터를 적극 활용하여, CLIP의 적용 범위를 다중 레이블 데이터셋 환경으로 확장하고자 한다. 이러한 맥락에서 CLIP 모델을 특정 데이터셋의 특성에 맞춰 효율적으로 적응시키기 위한 다양한 선행 연구들이 시도되어 왔다. 연구[2]의 CoOp은 텍스트 프롬프트를 학습 가능한 벡터로 변환하여 해당 데이터셋에 맞게 최적화하는 방식을 제안함으로써 few-shot 성능을 개선했다. 또한 연구[3]의 CLIP-Adapter는 특징 추출기 뒤에 경량화 된 레이어를 추가하여 적응력을 높였다. 하지만 이 두 연구 모두 여전히 역전파를 통한 추가적인 파라미터 학습을 요구하므로, 연산 자원이 제한되거나 즉각적인 적응이 필요한 환경에서는 효율성이 떨어진다는 단점이 있다. 본 논문에서는 훈련 없는 few-shot 전이 학습 기법인 Tip-Adapter[4]을 활용하여, CLIP의 사전 지식을 그대로 활용하면서도 few-shot 지식을 결합함으로써 Single Positive Label(SPL) 제약 조건 하에서의 다중 레이블 분류 성능을 분석했다. 이를 통해 복잡한 전처리나 추가 학습 없이도 기존 zero-shot 대비 향상된 성능을 보임을 확인하였으며 결과적으로 CLIP이 다중 객체 환경에 효과적으로 적응할 수 있음을 입증했다.

Ⅱ. 본론

2.1 Tip-Adapter

Tip-Adapter는 CLIP의 가중치를 고정한 채, 역전파 없이 few-shot 데이터만으로 새로운 도메인에 적응하는 훈련 없는 전이학습 기법이다. 본 방법론은 Support Set의 특징을 key로, 레이블 정보를 value로 하는 캐시(cache)를 구축하여 지식 베이스를 형성한다. 추론 시에는 쿼리 이미지와 캐시 내 key들 간의 코사인 유사도를 계산하여 few-shot 지식을 추출하고 이를 기존 CLIP의 zero-shot 예측 결과와 선형적으로 결합하여 최종 결과를 도출한다. 이는 별도의 반복 학습 과정이 불필요하여 연산 효율성이 뛰어나며, 소량의 데이터에서도 과적합 없이 안정적인 성능을 보장한다.[4]

2.2 데이터

본 논문에서는 SPL 환경에서의 성능 확인을 위해 다양한 객체가 혼재된 MS-COCO 2014 Dataset[5]을 사용했다. Support Set은 학습 데이터의 80개 클래스를 대상으로 1-shot부터 16-shot까지 각 K장의 이미지를 무작위로 추출하여 구성했다. 이때 캐시 모델이 객체의 존재 여부를 대조하여 학습할 수 있도록, 각 클래스 별로 해당 객체가 포함된 양성 이미지와 포함되지 않은 음성 이미지를 선별하여 Support Set을 구성하였다. 또한, 현실적인 제약 조건을 반영하기 위해 각 이미지에 할당된 다수의 양성 레이블 중 무작위로 선택된 단 하나의 레이블 정보만을 유지하고 나머지는 모두 제거하여 캐시 구축에 활용하였다. Query Set은 모델의 일반화 성능을 평가하기 위해 검증 데이터 전체를 사용했다. 추론 단계에서는 각 클래스에 대해 독립적으로 수행된 이진 분류 결과를 취

합하여 최종적인 다중 레이블 예측 벡터를 생성한다. 이후 모델의 실제 다중 객체 식별 능력을 검증하기 위해, 해당 결과 벡터를 원본 데이터셋이 보유한 다중 레이블 정답과 대조하여 성능을 측정하였다.

2.3 제안하는 방법

본 논문에서는 다중 객체가 혼재된 SPL 환경에 대응하기 위해, 클래스별 독립적인 캐시 기반의 프레임워크를 제안한다. 본 구조는 별도의 역전파 과정이 필요 없는 학습 비의존 방식으로, 그림1과 같이 세 가지 주요 단계를 거쳐 구동된다.

2.3.1. 특징 추출 및 캐시 구축. 먼저, 캐시 구축 단계에서는 SPL 데이터를 CLIP의 사전 학습된 비주얼 인코더에 입력하여 이미지의 전역 특징을 추출한다. 이후 각 클래스의 존재 여부를 수치화하기 위해 one-hot 인코딩을 적용하는데, 객체가 존재하는 양성 샘플은 [1,0]으로, 존재하지 않는 음성 샘플은 [0,1]로 처리한다. 그리고 이를 각각 캐시의 key, value로 구성하여, 모델이 양성과 음성 특징 중 어느 쪽과 더 유사한지 비교하였다. 이를 통해 각 객체의 존재 여부를 독립적으로 판단하는 이진 판별 구조를 구현하였다.

2.3.2. Zero-shot 및 Few-shot 지식 추출. 이후 추론 단계에서는 쿼리 이미지를 비주얼 인코더에 통과시켜 특징을 추출하고, 이를 구축된 캐시 모델 내의 이미지 특징(key)들과 비교하여 유사도를 계산한다. 이후 계산된 유사도 값에 value을 곱하여 쿼리 이미지와 유사한 샘플들의 정답 정보를 인출함으로써 해당 클래스에 대한 few-shot logit을 산출한다. 이와 동시에 CLIP의 텍스트 인코더를 통해 생성된 양성 및 음성 프롬프트 임베딩과 쿼리 이미지 간의 zero-shot logit을 구한다.

2.3.3. 지식 결합 및 최종 확률 도출. 최종적으로 CLIP의 사전 지식에 few-shot logit을 선형 결합하고, softmax을 적용하여 각 클래스의 존재 확률을 도출한다. 이후 총 80개의 클래스에 대해 위 과정을 반복 수행함으로써 각 객체의 존재 여부를 독립적으로 판별하고, 이를 취합하여 최종적인 다중 레이블 분류 결과를 생성한다.

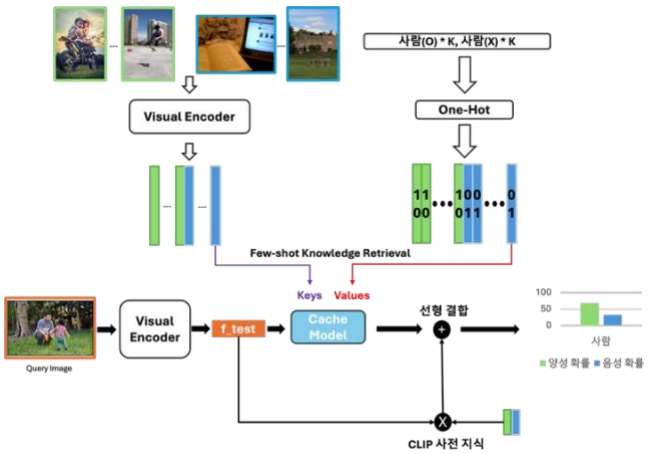
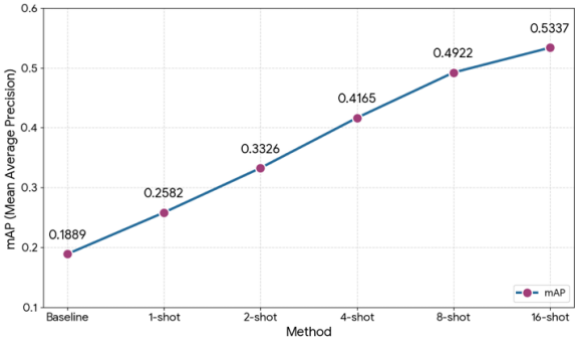


그림1. 제안하는 프레임워크의 전체 구조

2.4 실험 결과 및 분석

본 논문에서는 다중 레이블 분류 성능을 정량적으로 측정하기 위해 mAP(mean Average Precision)를 지표로 사용했다. mAP는 각 클래스의 정밀도-재현율 곡선 아래 면적을 평균한 값으로, 다중 객체가 혼재된 환경에서 모델의 분류 정확도를 종합적으로 평가하는 데 적합하다. 제안하는 캐시 프레임워크를 적용하여 shot 수(K=1, 2, 4, 8, 16) 증가에 따른 성능 변화를 측정하고, 이를 기존 zero-shot 성능(Baseline)과 비교하여 그래프[1]에 정리했다. 실험 결과, 학습 데이터에 단 하나의 양성 레이블만 존재하는 SPL의 열악한 환경임에도 불구하고, few-shot 데이터가 제공됨에 따라 성능이 뚜렷하게 향상되었다. 이를 통해 추가 파라미터 학습 없이 적은 수의 SPL 이미지만으로도 다중 객체 분류 성능을 효과적으로 개선할 수 있음을 확인했다.



그래프1. Shot 수에 따른 mAP 성능 분석

Ⅲ. 결론

본 논문에서는 Tip-Adapter를 활용하여 CLIP의 사전 지식을 유지한 채 SPL 제약 조건 하에서의 다중 레이블 분류 효율성을 입증했다. 막대한 레이블링 비용 문제를 해결하기 위해, 구축이 용이한 단일 양성 레이블 데이터와 클래스별 독립 이진 판단 구조를 결합했다. 실험 결과, 역전파를 통한 추가 학습 없이 few-shot 데이터의 추가만으로도 zero-shot 대비 유의미한 성능 향상을 달성했다. 이는 고비용의 전체 레이블링이나 모델 재학습 없이도 대규모 비전-언어 모델을 다중 객체 환경에 효과적으로 적응시킬 수 있음을 시사한다.

ACKNOWLEDGMENT

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 정보통신방송혁신인재양성(메타버스융합대학원) 사업 연구 결과로 수행되었습니다. (IITP-2026-RS-2023-00254529)

본 연구는 2025년도 정부(과학기술통신부)의 재원으로 국가과학기술연구회 글로벌 TOP 전략연구단 지원사업 (No.GTL25101-301)의 지원을 받아 수행되었습니다.

참 고 문 헌

[1] A. Radford, et al. “Learning Transferable Visual Models From Natural Language Supervision,” in ICML, 2021.

[2] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, “Learning to Prompt for vision-language models,” International Journal of Computer Vision (IJCV), vol. 130, no. 9, pp. 2337-2348, 2022.

[3] P. Gao, et al. “CLIP-Adapter: Better vision-language models with feature adapters,” arXiv preprint arXiv:2110.04544, 2021.

[4] R. Zhang, et al. “Tip-Adapter: Training-free CLIP-Adapter for Better Vision-Language Modeling,” in European Conference on Computer Vision (ECCV), 2022.

[5] T. -Y. Lin, et al. “Microsoft COCO: Common objects in context,” in European Conference on Computer Vision (ECCV), 2014.