시각-언어 모델을 위한 이미지 개별 프롬프트 학습

이인수, 심병효
서울대학교

{insu301, bshim}@snu.ac.kr

Instance-Wise Prompt Learning for Vision-Language Models

Lee In Su, Shim Byong Hyo
Seoul National Univ.

요 약

본 논문은 시각-언어 모델(Vision-Language Models)을 이용한 이미지 분류에서 클래스 내 변동성 문제를 해결하기 위해 이미지별 프롬프트 학습을 제안한다. 기존 연구는 각 클래스마다 사전 정의된 클래스별 프롬프트를 사용하는데, 이는 클래스 내 모든 이미지의 변동성을 포착하지 못할 수 있다. 해당 기법은 각 이미지마다 고유한 프롬프트를 할당하여 이미지 간 구별을 강화하고 전체 분류 정확도를 향상시킨다.

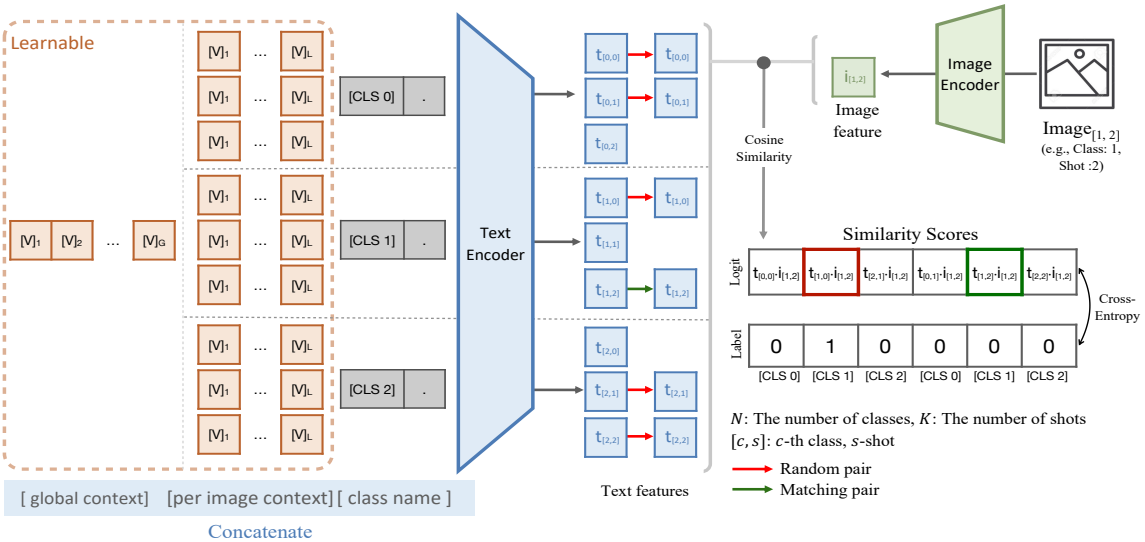
Ⅰ. 서 론

최근 사전 훈련된 시각-언어 모델은 인간의 언어를 사용해 시각 하위 작업을 수행 할 수 있게 하였다. 그 중 CLIP [1] 모델은 텍스트와 이미지 간의 공유 임베딩 공간을 대조 학습을 통해 학습한다. 이는 특정 이미지에 가장 적합한 텍스트를 식별할 수 있게 한다. 시각-언어 모델을 사용한 이미지 분류를 위해서는 각 클래스를 대표하는 단일 텍스트, 즉 프롬프트를 정의하는 것이 필수적이다. 그러나 기존의 고정된 프롬프트는 해당 클래스를 적절히 표현하지 못한다. 이를 해결하기 위해, CoOp 은 [2] 학습 가능한 프롬프트를 도입하고 각 클래스를 가장 잘 대표하는 단일 프롬프트를 찾는다. 그러나 각 클래스마다 단일 대표 프롬프트를 찾는 것은 최적이 아닐 수 있다. 같은 클래스에 속하더라도 이미지 각각은 차이가 있으며, 이러한 차이들을 하나의 프롬프트로 포착하는 것은 어려운 일이다.

본 연구는 이미지별 프롬프트 학습을 제안한다. 훈련 데이터의 각 이미지에 다른 프롬프트를 할당하여, 동일 클래스 내 다른 이미지를 잘 구별함과 동시에 동일 클래스의 공통 특성을 잘 포착하는 것을 목표로 한다.

Ⅱ. 본론

본 논문은 이미지별 프롬프트 학습을 제안한다. 구체적으로 이미지별 프롬프트는 다음과 같이 세 가지 요소로 구성된다: 전체적 맥락 벡터, 이미지별 맥락 벡터, 그리고 클래스 이름. 전체적인 맥락 벡터는 모든 클래스와 이미지에 걸쳐 공유되는 학습 가능한 벡터로, 전체 데이터셋에 관련된 공통 특징이나 맥락을 파악한다. 이미지별 맥락 벡터는 특정 클래스의 특정 이미지에



관련된 특징과 맥락을 파악한다.

훈련 시 각 이미지에 대한 프롬프트 선택 과정은 다음과 같다: 각 클래스마다 2 개의 프롬프트를 무작위로 선택하되, 이미지와 매칭되는 프롬프트 하나를 결정적으로 선택한다. 선택된 프롬프트와 이미지 사이의 코사인 유사도를 계산 한 후, 크로스-엔트로피 로스를 이용해 훈련을 진행한다.

훈련 후 테스트 이미지가 입력 될 때, 해당 이미지와 학습된 모든 프롬프트 간의 코사인 유사도를 계산하고 각 클래스마다 유사성이 가장 높은 프롬프트를 4 개 선택한다. 선택된 프롬프트를 평균을 내어 각 클래스를 대표하는 하나의 프롬프트를 생성하고, 이를 사용하여 테스트 이미지와 각 프롬프트 간의 코사인 유사도를 계산한다. 최종적으로 가장 높은 코사인 유사도를 보이는 클래스를 테스트 이미지의 클래스로 지정한다. 이러한 과정을 통해, 본 연구는 더 정확하고 효율적인 이미지 분류를 가능하게 한다.

클래스별 이미지 4 장과, ResNet-50 [3] 을 이용하여 다양한 데이터셋에서 실험한 결과, 기존의 모델들에 비해 성능이 향상된 것을 확인 할 수 있다.

Method	Backbone	N-shot	Flowers102	Caltech101	Pets	Aircraft
ZS CLIP			66.14	86.29	85.77	17.28
LP CLIP			83.89	84.89	57.23	24.62
CoOp	ResNet-50	4-shot	86.20	89.55	86.70	21.87
LFA			87.10	91.04	87.80	22.31
Ours			88.20	90.56	88.21	23.46

Ⅲ. 결론

본 연구는 각 이미지에 특화된 프롬프트를 할당하는 새로운 이미지별 프롬프트 학습 방법을 제안하였다. 이 방법은 동일 클래스 내 다른 이미지들을 효과적으로 구별하면서 동시에 클래스의 공통적인 특성을 정확히 포착한다. 초기 실험을 통해 본 방법이 기존 모델에 비해 성능이 향상됨을 확인하였다. 이 기법을 발전시켜 더욱 향상된 성능을 가진 이미지 분류 모델을 개발할 수 있을 것으로 기대한다.

ACKNOWLEDGMENT

이 논문은 2022 년도 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2022M3C1A3099336).

참 고 문 헌

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning, pages 8748- 8763. PMLR, 2021.

[2] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. International Journal of Computer Vision, 130(9):2337-2348, 2022.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. corr abs/1512.03385 (2015), 2015.