제로 샷 학습을 통한 얼굴 표정 이미지와 텍스트의 관계 추론

Inference of the relationship between facial expression images and text through zero-shot learning

한채림 이덕우

Department of Computer Engineering, Keimyung University

cozyriming@gmail.com dwoolee@kmu.ac.kr

1. Abstract

This paper employs the CLIP model, which consists of Vision Transformer (ViT) to process images and Text Transformer to process texts, based on the Transformer architecture. With this model, facial expressions, traditionally categorized into seven expressions, are learned through textual sentences. Compared to conventional methods of facial expression recognition, this approach enables recognition and classification of diverse and complex facial expressions through descriptions of facial expressions using textual sentences.

2. 제로 샷 학습 (Zero-Shot Learning)

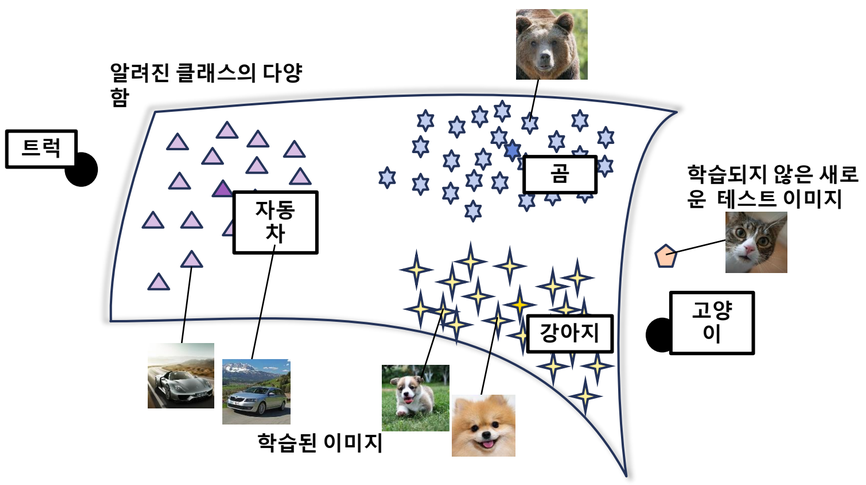


그림 1. 제로 샷 학습 구조

3. CLIP 모델(Contrastive Language-Image Pre-training model)

CLIP 모델은 이미지와 텍스트 간의 관계를 이해하도록 설계된 모델이다. 대조 학습(Contrastive Learning) 방식을 따르며, 이는 모델이 주어진 이미지에 가장 잘 맞는 텍스트를 선택하거나, 반대로 주어진 텍스트에 가장 잘 맞는 이미지를 선택하도록 학습한다. 이 과정을 통해 모델은 이미지와 텍스트 간의 복잡한 관계를 파악한다. 또한 본 논문에서 사용하는 CLIP 모델은 전체 문장과 문장이 설명하는 이미지 사이의 관계를 학습한다. '강아지', '자동차', '곰' 등과 같은 단일 클래스가 아닌 문장 전체에 대해 훈련이 된다. 따라서 위 모델을 이용하여 문장이 주어지면 해당 문장에서 가장 관련성이 높은 이미지를 검색하거나 이미지 데이터와 텍스트 데이터에 대해 학습할 때 분류기 역할도 가능하게 한다. CLIP 모델의 이러한 점들을 이용하여 얼굴 표정 이미지와 한 문장의 얼굴을 설명하는 텍스트를 CLIP 모델을 사용하여 이미지와 텍스트가 유사도가 높게 측정되도록 학습한다.

4. Experiments

본 논문에서 제로 샷 학습을 위해서 실험 환경은 Google의 Colab 환경에서 pytorch를 사용하고, 학습을 위한 GPU로는 Colab에서 지원하는 T4 GPU를 사용하였다. 배치 크기는 16, $epochs$는 50번과 학습률은10^{-4}로 설정한다. 이미지 인코딩의 학습률은 e^{-4}를 사용하고, 텍스트 인코딩에서는 e^{-5}로 학습을 진행하였다. 이미지의 사이즈는 224×224(pixels)를 따르고, 초기의 사전 학습된 모델로는 resnet 50을 따른다. 이미지 모델로는 ViT 모델을 사용하였고, 텍스트 모델로는 DistilBERT모델을 따른다. 학습에 사용된 이미지는 기본 표정 7가지에 대해서 각 4,000장씩 사용된다. 학습에 사용된 텍스트 라벨은 아래와 같다. 아래의 문장을 텍스트 인코더에 들어가면 DistilBERT 토크나이저를 사용하여 문장(캡션)을 토큰화한 다음 토큰 ID와 어텐션 마스크가 되어 학습이 진행된다.

0	neutral	"An expression of calm and neutrality, with a neutral mouth and no particular indication of emotion. The eyebrows are usually not raised or furrowed"
1	happiness	"An expression of contentment and pleasure, with a smile and the corners of the mouth turned up, often accompanied by crinkling around the eyes. The face may appear relaxed and at ease"
2	sadness	"An expression of sadness and sorrow, with a downturned mouth or frown, and sometimes tears or a tightness around the eyes. The face may appear physically withdrawn or resigned."
3	surprise	"An expression of shock and astonishment, with wide-open eyes and raised eyebrows, sometimes accompanied by a gasp or an open mouth"
4	fear	"An expression of tension and withdrawal, with wide-open eyes, raised eyebrows, and a slightly open mouth. The face may appear physically tense or frozen in fear"
5	disgust	"An expression of repulsion and displeasure, with a raised upper lip, a scrunched nose, and a downturned mouth"
6	anger	"A facial expression showing irritation and unrest, with a wrinkled forehead, narrowed eyes, and tight lips or a frown"
7	contempt	"An expression of disdain and superiority, with a slight smirk or sneer, often accompanied by a raised eyebrow or a lopsided smile"

표 1. 학습에 사용된 문장

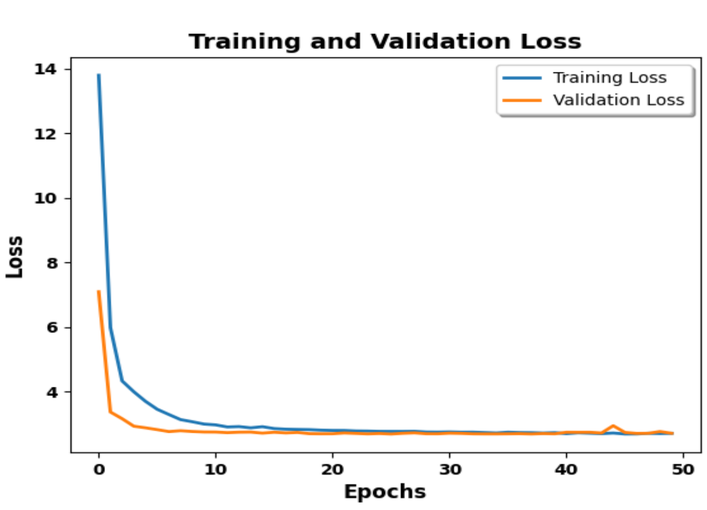
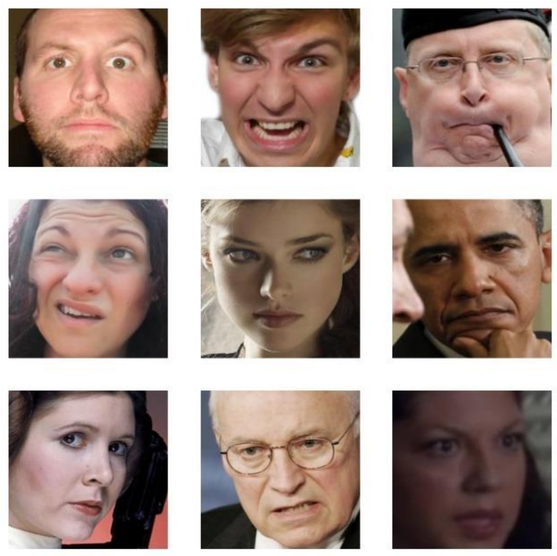


그림 2. Epochs 50번에 대한 Loss 결과



query	An expression of disdain and superiority.

그림 3. 이미지와 텍스트 쌍 매칭 결과

그림 2는 $epochs$ 50번의 학습 과정에서 학습 데이터와 검증 데이터의 손실함수를 나타낸다. 이미지와 텍스트를 각각 256차원의 공간에 투영시킨다. 이미지 임베딩과 텍스트 임베딩 된 두 개의 벡터 그룹이 서로 얼마나 유사한지를 본다. 유사도를 판단하는 방법으로는 선형 대수학에서 쓰이는 내적을 사용한다. 이미지 벡터와 텍스트 벡터를 곱하여 최종 숫자가 크면 비슷하고 작으면 다른 것을 의미한다. 손실 함수 모델의 성능을 평가하는 데 사용하며, 예측값과 실제값 사이의 차이를 수치화한다. 손실 함수로는 교차 엔트로피를 사용한다. 교차 엔트로피 값이 크다는 것은 모델의 예측이 실제 레이블과 크게 다르다는 것을 의미하며, 모델의 성능이 좋지 않고 예측에 대한 불확실성이 높다는 것을 나타낸다. 따라서 손실 함수 그래프의 기울기가 줄어들수록 이미지에 대한 텍스트가 유사함을 알 수 있다. 처음 1 $epochs$에 학습 $loss$ 값은 13.8이었으나, 2 $epochs$에서는 학습 $loss$ 값이 5.98로 현저하게 감소함을 알 수 있다. 학습이 끝날 시점 인 50 $epochs$ 에서는 최종적으로 학습 $loss$와 검증 $loss$ 모두 2.71로 처음과 비교하여 11.09 감소하였다. 아래의 그림 3은 마지막으로 학습 후 최종 테스트 이미지에 대한 결과이다. 텍스트 쿼리로 "An expression of disdain and superiority."를 입력하고, 가장 유사한 이미지 9개에 대한 출력을 나타낸다. 쿼리 문장에서 "disdain"이나 "superiority"를 직역하면 "경멸감", "우월"이라는 뜻이다. 문장의 의미와 이미지가 유사하게 출력되고 있음을 알 수 있다.

5. Conclusion

기존에 널리 분류되어 온 7가지 기본적인 얼굴 표정을 단어가 아닌 문장으로 묘사하여 더 다양하고 복잡한 얼굴 표정의 인식 및분류를 진행해 보았다. 이러한 접근은 기존의 얼굴 표정 인식 방법들과 비교했을 때 더 포괄적이고 세밀한 분석이 가능하게 하며, 복잡한 인간 감정의 이해에 있어 중요한 발전을 이룰 수 있다고 생각한다. 본 연구를 통해 얻은 결과는 향후 초거대 인공지능 연구에 있어 중요한 기여를 할 것으로 기대된다. 특히, 감정과 같은 비언어적인 요소를 더욱 정교하게 인식하고 분석할 수 있는 모델의 개발에 있어서도 앞선 연구의 결과가 활용될 수 있을 것이다. 앞으로의 연구에서는 CLIP 모델의 성능을 더욱 향상시키고, 다양한 비언어적 요소를 인식할 수 있는 모델의 범용성을 확장하는 방향으로 나아가고자 한다. 이를 위해 더욱 다양하고 방대한 데이터셋을 활용한 학습과, 모델의 구조적 개선에 대한 연구가 필요할 것이다. 또한, 인공지능이 인간의 감정과 비언어적 신호를 보다 정확하게 이해하고, 의사소통을 보다 깊이 있게 이해하고, 이를 바탕으로 한 자연스러운 인간-인공지능 간 상호작용의 실현을 목표로 한다.