제로샷 학습을 통한 얼굴 표정 이미지와 텍스트의 관계 이해

한채림, 이덕우
계명대학교 컴퓨터공학과
e-mail : cozyriming@gmail.com, dwoolee@kmu.ac.kr

Understanding the relationship between facial expression images and text through Zero-Shot Learning

Chaerim Han, Deokwoo Lee

Department of Computer Engineering
Keimyung University

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.

I. 서론

최근 심층신경망 모델을 활용한 심층학습 기술이 발전함에 따라 인공지능 기술 및 적용 범위 또한 확대되고 있다. 특히 트랜스포머 아키텍처를 기반으로 하는 대규모 언어모델들은 억 단위의 매개 변수를 학습하며, 자연어 처리, 컴퓨터 비전, 음성 인식 등 많은 분야에서 우수한 성능을 보이고 있다. 그러나

기존의 모델들은 이미지나 텍스트 그리고 오디오 데이터를 인코더 디코더 구조로 한 방향의 데이터 흐름을 처리하는데 집중한다. 그러나, 트랜스포머 모델은 이러한 구조를 확장한다. 트랜스 포머는 “Attention is All You Need”라는 논문에서 처음 소개되었으며, 주의 집중 메커니즘(Attention Mechanism)으로 구성되어 있다.[1] 주의 집중 메커니즘은 ‘기계 번역’ 수행에서 비롯되었으며, 디코더에서 출력 단어를 예측하는 매 시점마다 인코더의 입력 시퀀스를 다시 참고한다.[2] 이 구조는 모델이 입력 데이터의 다양한 부분에 주목할 수 있게 해주어, 사람의 감정 및 표정 인식과 같은 복잡한 비언어적 인지 기능에 대한 연구를 효과적으로 학습할 수 있게 한다. 이를 위해 본 연구에서는 이미지 캡션 쌍 데이터를 활용한 제로샷 학습 기법을 사용하며 트랜스포머 모델 구조로 얼굴 표정 인식을 제안한다. 모델이 미리 학습한 적 없는 새로운 객체나 개념을 인식하고 분류할 수 있도록 한다. 이 방식을 통해 기본 7가지 분류의 단어가 아닌, 얼굴 표정 묘사가 가능한 문장 통해서 다양하고 복잡한 표정 인식과 분류를 진행해본다.

II. 본론

2.1 제로샷 학습(Zero-Shot Learning, ZSL)

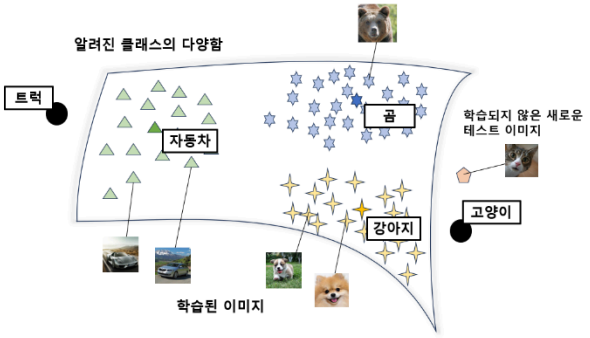


그림 1. 제로샷 학습 구조

제로샷 학습(Zero-Shot Learning)은 기계 학습의 한 분야로, 모델이 학습 과정에서 보지 못한 새로운 클래스를 인식하거나

분류할 수 있게 한다.[3] 이는 특히 데이터가 부족하거나 특정 작업에 대한 레이블이 지정된 데이터를 얻기 어려운 경우에 유용하다.

그림 1과 같이 모델이 '강아지', '곰'과 같은 동물 클래스를 학습한다. 이후 제로샷 학습을 통해 모델은 '고양이'와 같은 학습 과정에서 본 적 없는 새로운 클래스를 인식할 수 있게 된다. 이는 '강아지'가 공유하는 특성인 '네 발로 걷는다, 털이 있다.' 등을 기반으로 한다. 또한 그림 1과 같이 데이터의 클래스를 고차원 공간에 표현하는 방법을 클래스 임베딩이라고 한다. '강아지'와 '곰'은 동물이라는 공통점으로 임베딩 공간에서 가까운 반면, '자동차'는 먼 거리에 위치한다. 이처럼 클래스 임베딩은 본 적 없는 클래스와 본 적 있는 클래스를 연결하는 데 핵심 구성 요소가 되며, 데이터의 복잡한 구조나 관계를 이해하고, 데이터 포인터들 사이의 유사성을 계산한다.[4] 이미지와 클래스 임베딩을 비교하기 위해 대부분의 제로샷 학습 방법은 이미지와 클래스가 의미론적 의미에 따라 매핑되는 시맨틱 공간을 도입한다. [5, 6, 7]

2.2 CLIP 모델(Contrastive Language-Image Pre-training model)

CLIP 모델은 OpenAI가 개발한 고유한 접근 방식을 사용하여 이미지와 텍스트 간의 관계를 이해하도록 설계된 모델이다.[8] 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.

Ⅲ. 실험 및 분석

본 논문에서 제로샷 학습을 위해서 실험 환경은 Google의 Colab환경에서 pytorch를 사용하고, 학습을 위한 GPU로는 Colab에서 지원하는 T4 GPU를 사용하였다. 배치 사이즈는 16, 에폭 횟수는 50번과 학습률은 10^{-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 resnet50을 따른다. 이미지 모델로는 ViT 모델을 사용하였고, 텍스트 모델로는 DistilBERT모델을 따른다. 학습에 사용된 이미지는 기본 표정 7가지에 대해서 각 4000장씩 사용된다. 학습에 사용된 텍스트 라벨은 아래와 같다. 아래의 문장을 텍스트 인코더에 들어가면 DistilBERT 토크나이저를 사용하여 문장(캡션)을 토큰화 한 다음 토큰ID와 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

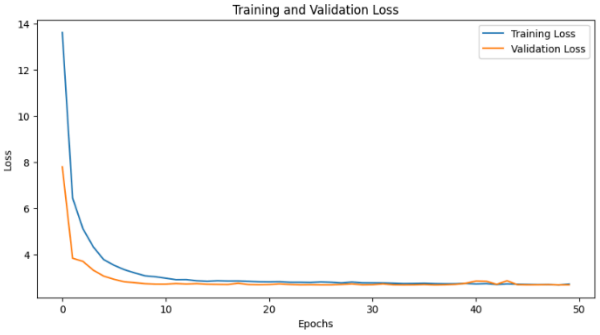
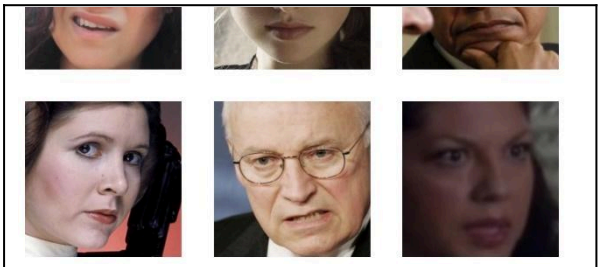


그림 2. Epochs 50에 대한 Loss 결과

그림 2는 50번의 학습과정에서 학습 데이터와 검증 데이터의 손실함수를 나타낸다. 이미지와 텍스트를 각각 256차원의 공간에 투영시킨다. 이미지 임베딩과 텍스트 임베딩된 두 개의 벡터 그룹이 서로 얼마나 유사한지를 본다. 유사도를 판단하는 방법으로는 선형 대수학에서 쓰이는 내적본다. 이미지 벡터와 텍스트 벡터를 곱하여 최종 숫자가 크면 비슷하고 작으면 다른 것을 의미한다. 손실 함수모델의 성능을 평가하는데 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 $1epochs$에 학습 $loss$ 값은 13.6이었으나, $2epochs$ 에서는 학습 $loss$ 값이 7.79로 현저하게 감소함을 알 수 있다. 학습이 끝날 시점인 $50epochs$에서는 최종적으로 학습 $loss$가 2.72로 검증 $loss$ 는 2.68로 처음과 비교하여 감소하였다. 아래의 그림 3는 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.



query	An expression of disdain and superiority.

그림 3. 이미지와 텍스트 쌍 매칭 결과

Ⅳ. 결론 및 향후 연구 방향

기존에 널리 분류되어 온 7가지 기본적인 얼굴 표정을 단어가 아닌 문장으로 묘사하여 더 다양하고 복잡한 얼굴 표정의 인식 및 분류를 진행해보았다. 이러한 접근은 기존의 얼굴 표정 인식 방법들과 비교했을 때 더욱 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.

참고문헌

[1] C. Subakan, M. Ravanelli, S. Cornell, M. Bronzi and J. Zhong, "Attention Is All You Need In Speech Separation," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021:21-25

[2] Luong, T., Pham, H., & Manning, C.D. (2015). Effective Approaches to Attention-based Neural Machine Translation. ArXiv, abs/1508.04025:3-5

[3] Romera-Paredes B, Torr P. An embarrassingly simple approach to zero-shot learning. Proceedings of the 32nd International Conference on Machine Learning, PMLR. 2015:1.

[4] Y. Xian, C. H. Lampert, B. Schiele and Z. Akata, "Zero-Shot Learning—A Comprehensive Evaluation of the Good, the Bad and the Ugly," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 9:2

[5] E. Kodirov, T. Xiang and S. Gong, "Semantic Autoencoder for Zero-Shot Learning," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 4447-4456, doi: 10.1109/CVPR.2017.473.

[6] Fu, Y., Hospedales, T.M., Xiang, T., Gong, S.: Transductive Multi-View Zero-Shot Learning. IEEE T-PAMI 37(11), 2332–2345 (2015)

[7] Y. Zhang, B. Gong and M. Shah, "Fast Zero-Shot Image Tagging," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 5985-5994, doi: 10.1109/CVPR.2016.644.

[8] Conde, Marcos V, Turgutlu, Kerem. CLIP-Art: Contrastive Pre-Training for Fine-Grained Art Classification. Proceedings of the IEEE/CVF

Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. 2021:2