멀티모달 기반 얼굴 및 음성 표현 정렬을 통한 우울 관련 비언어적 신호 탐지에 관한 연구

최재민, 황채아, 윤수연*
국민대학교, *국민대학교
kiy9514@kookmin.ac.kr, hcaopen@kookmin.ac.kr, *1104py@kookmin.ac.kr

A Study on Duo-Related Language Signals for Multimodal-Based Facial Expression Representation

Jae Min Choi, Chae a Hwang, Soo Yun Yoon*
Kookmin Univ., *Kookmin Univ.

요 약

본 논문은 임상 라벨에 의존하지 않고, 얼굴 표정 기반 감정 표현을 학습을 통해 우울 관련 비언어적 신호 분석을 위한 기반 모델을 구축하는 것을 목표로 한다. AI Hub 감정 복합 영상 데이터셋을 활용하여 감정 텍스트와 정적 얼굴 이미지를 정렬한 감정 특화 CLIP 모델을 제안한다. 학습된 모델은 zero-shot 감정 분류 방식으로 평가하며, 7개 감정 클래스에 대해 56.28%의 정확도를 기록하였다. 중립과 슬픔 감정에서 비교적 안정적인 분류 성능을 보였다. 본 논문은 임상 라벨이 의존하지 않고도 얼굴 표정 기반 감정 표현 학습이 가능함을 보인다.

Ⅰ. 서 론

우울증은 전 세계적으로 흔한 정신건강 질환으로 삶의 질 저하와 사회적 및 경제적 부담을 유발하고 있다. 세계보건기구(WHO)에 따르면 우울증은 자살과 밀접하게 관련되어 있으며 조기 발견과 적절한 개입이 이루어질 때 자살 예방 및 우울감 완화가 가능하다고 한다.[1] 그러나 일상생활에서는 우울증 증상 인지하지 못하거나 인지하더라도 사회적 낙인으로 인해 전문가의 도움을 받는 시점이 지연되어 조기 발견에 어려움이 존재한다고 보고되고 있다.[2]

우울 관련 위험 신호는 얼굴 표정, 대화, 행동과 같은 언어적·비언어적 표현에 복합적으로 나타나는 것으로 보고되고 있다.[3],[4] 이러한 특성으로 인해 단일 모델만으로는 신뢰도 측면에서 한계를 가진다. 또한, 실제 현장에서는 PHQ-9와 같은 임상 라벨 확보나 접근성 제약으로 인해 승인이 어렵다. 이런 점으로 인해 본 연구는 멀티모달 표현을 정렬하기 위해 감정 이미지와 텍스트를 정렬하여 우울 관련 비언어적 신호를 탐지하는 것으로 연구를 설계하려 한다.

이에 본 연구는 PHQ-9 데이터로 직접적으로 수행하는 대신, 우울 관련 비언어적 신호를 탐지하기 위한 선행 단계로 얼굴 표정에 나타나는 감정 표현을 학습하는 것을 연구 목표로 설정한다. 이를 위해 정적 얼굴 이미지와 감정 텍스트의 의미를 정렬하여 감정 표현에 민감한 시각 모델을 구축하고, 해당 표현을 zero-shot 검증을 통해 감정 분류 성능 및 예측 분포 분석을 진행한다. 추후 학습된 표현 모델은 영상 데이터에서 얼굴, 음성, 텍스트 정보를 실시간으로 정렬하는 우울 관련 신호 탐지 연구로 확장될 수 있는 기반으로 활용할 예정이다.

Ⅱ. 본 론

2.1. AI 기반 안면인식 기술

AI 기반 안면인식 기술은 영상에서 얼굴을 검출하고 특징을 추출하여 개인 식별 및 감정 등을 분석하는 기술이다. 현재는 합성곱신경망(CNN) 기반 기법이 표준으로 자리 잡았다. CNN 기반 안면인식 기술은 일반적으로 검출, 특징 추출, 분류의 기술들로 구성된다.

표정 인식 분야의 최근 연구들은 CNN, ResNet, EfficientNet 등 심층 신경망을 활용하여 정적 이미지뿐만 아닌 영상 시퀀스 기반 표정 변화를 모델링하는 방법으로 활발히 연구되고 있다.[5] 본 논문에서는 이러한 기술들에 아이디어를 얻어 얼굴 표정의 시각적 특징을 추출 및 정량화한다.

2.2. 멀티모달 우울 관련 연구 동향

우울과 관련된 비언어적 신호는 앞에서 서술했듯이 다양한 형태로 관찰되고 있다. 이에 따라 기존 연구들은 얼굴 영상 또는 음성 신호 중 하나만을 활용하는 접근의 한계를 지적하고, 여러 모달을 결합하여 우울 관련 신호를 정확하게 탐지하려는 방향으로 연구 중이다.

실제로 한 연구에서는 음성 및 영상 데이터에서 딥러닝 기반 특징을 결합해 우울증 탐지의 성능을 향상시키는 구조를 발표했으며, 얼굴 표정과 발화 기반 신호를 조합하여 우울 관련 특성을 반영할 수 있음을 증명하였다.[6]

또한, Transformer를 활용한 연구도 활발하다. IMDD-Net은 TimeSformer를 이용하여 비디오 시공간 정보를 BERT 기반 텍스트 인코더에 결합하여 영상, 음성, 텍스트 간 상호작용을 통합하는 접근을 제안하였다.[7] 그러나 기존 멀티모달 및 Transformer 기반 연구들은 대부분 PHQ-9과 같은 임상 라벨에 의존한 학습은 전제로 한다. 따라서 임상 라벨에 의존하지 않고 성립 가능한 우울 관련 비언어적 신호 분석을 위한 실험적 접근을 제안하고자 한다.

Ⅲ. 실험방법

본 연구에서는 우울과 연관된 비언어적 신호 분석을 위한 선행 단계로서, 정적 얼굴을 기반으로 감정 표현에 특화된 CLIP 모델 학습을 수행한다. 이를 통해 PHQ-9 데이터에 의존하지 않고, 얼굴 표정과 표정에 매칭되는 감정 표현을 언어적 공간과 정렬된 시각 표현을 학습한다.

3.1. 실험 개요

얼굴 이미지와 감정 텍스트 간의 의미를 정렬하여, 감정 표현에 민감한 시각 임베딩을 학습하는 것을 목표로 한다. 이를 위해 CLIP 모델을 기반으로 감정 특화 파인튜닝을 수행하였으며 학습된 모델의 성능은 이후 실험 및 성능 평가 단계에서 zero-shot 감정 분류 방식으로 분석하였다.

3.2. 데이터 구성

- 원본 데이터

원본 데이터는 AI Hub의 감정 복합 영상 데이터를 사용하였다. 해당 데이터는 여러 사람의 표정 이미지 파일과 JSON 형식으로 구성되며 JSON에는 어노테이션과 얼굴 bounding box 좌표, 감정 라벨 정보가 포함되어 있다. 얼굴 영역에 집중하기 위해 bounding box를 이용해 얼굴 영역을 직접 크롭하여 학습 데이터를 구성한다.

사용되는 주요 어노테이션 정보는 다음과 같다.

• 얼굴 bounding box: $(x_{\min}, y_{\min}, x_{\max}, y_{\max})$

• 감정 라벨: 기쁨, 당황, 분노, 불안, 상처, 슬픔, 중립.

- 얼굴 크롭 및 전처리

각 원본 이미지에서 얼굴 bounding box를 이용해 얼굴 영역을 크롭하고, 이를 CLIP 입력 크기에 맞게 리사이즈한다(예: 224x224).

원본 레코드 중 112,635장의 얼굴 이미지를 학습 데이터로 사용하였다.

3.3. 감정 텍스트 프롬프트 설계

CLIP 모델은 이미지와 자연어 문장 간의 의미 정렬을 기반으로 학습되므로, 감정 라벨을 단일 단어가 아닌 문장 형태의 텍스트 프롬프트로 변환하였다. 이는 문장 기반 프롬프트가 시각-언어 정렬 안정성을 향상시킨다는 것으로 알려져 있으며[8], 본 연구에서는 아래 Table 1과 같은 5종의 고정 템플릿 프롬프트를 설계하여 사용한다.

Table 1. CLIP 학습에 사용된 감정 텍스트 템플릿

Prompt ID	Text Template
P1	이 사람의 표정은 {감정}이다.
P2	이 사람은 {감정} 감정을 느끼는 것 같다.
P3	표정이 {감정}로 보인다.
P4	이 얼굴은 {감정}에 가깝다.
P5	이 사람은 지금 {감정}한 상태처럼 보인다.

3.4. 감정 특화 CLIP 학습

본 연구에서는 OpenAI에서 사전 학습된 CLIP ViT-B/32 모델을 초기 가중치로 사용하여 파인튜닝을 수행하였다. 학습은 이미지 인코더와 텍스트 인코더를 동시에 업데이트하는 방식으로 진행되었으며, 동일한 감정의 이미지 - 텍스트 쌍은 임베딩 공간에서 가깝게, 서로 다른 감정 쌍은 멀어지도록 하는 대비 학습(contrastive learning) 목표를 따른다.

Ⅳ. 실험 및 성능 평가

본 논문에서는 감정 특화 CLIP 모델의 성능을 정량적·정성적으로 평가한다. 평가는 별도의 분류기를 학습하지 않는 zero-shot 감정 분류 방식으로 수행되었으며, 감정 간 혼동 양상과 예측 분포를 중심으로 모델의 특성을 분석한다.

4.1 평가 설정

- 검증 데이터 구성

검증에는 학습에 사용되지 않은 얼굴 이미지 데이터만을 사용하였다. 각 감정 클래스는 약 7,400장의 이미지를 포함하고 있으며 검증 데이터 규모는 약 52,100장을 사용한다.

- Zero-shot 평가 방식

감정 분류를 위해 별도의 classifier head를 추가하지 않고, CLIP의 이미지-텍스트 임베딩 유사도를 이용하여 Zero-shot 평가를 수행하였다. 각 얼굴 이미지를 이미지 인코더에 입력하여 임베딩을 추출한 뒤, 7개 감정 클래스에 대응하는 텍스트 프롬프트 임베딩과의 cosine similarity를 계산하였다. 이때 각 감정은 단일 문장이 아닌 5종의 프롬프트 템플릿으로 표현되며, 하나의 이미지에 대해 다수 프롬프트와의 유사도를 종합하여 최종 예측 감정을 결정하였다.

4.2 정량적 성능 평가

7개 감정 클래스에 대한 zero-shot 평가 결과로는 전체 검증 데이터(52,107 samples) 기준 Top-1 정확도 56.28%를 기록하였다. 이는 정적 얼굴 이미지 정보만을 사용하고, 이미지-텍스트만 학습했음을 고려할 때 의미 있는 성능으로 판단된다.

4.3 예측 결과 및 혼동 행렬 분석

모델의 zero-shot 감정 분류 결과를 클래스별로 분석한 결과, 중립과 슬픔 감정이 상대적으로 높은 예측 빈도를 보였다. 이는 해당 감정들이 표정에서 비교적 일관된 시각적 특징을 가지며, CLIP 기반 모델이 이를 안정적으로 포착하고 있음을 시사한다.

반면, 불안, 상처, 당황과 같은 감정은 예측 빈도가 낮았으며, 표정만으로 명확히 구분하기 어려운 정서 상태로 인해 혼동 가능성이 높은 것으로 나타났다. 실제 데이터 분포와 예측 분포 간의 차이는 모델이 단순히 클래스 빈도를 모방하기보다, 시각적 특징에 기반한 의미적 판단을 수행하고 있음을 보여준다.

Table 2. zero-shot 감정 분류의 혼동 행렬 결과

	기쁨	당황	분노	불안	상처	슬픔	중립
기쁨	6592	334	65	116	47	291	308
당황	338	5202	242	469	280	488	610
분노	389	793	3082	615	376	873	492
불안	268	1207	333	1311	535	1138	828
상처	191	831	273	516	658	1499	743
슬픔	355	562	336	671	831	5501	811
중립	425	995	210	532	530	1074	6981

혼동 행렬 분석에서도 유사한 경향이 확인되었으며, 중립 감정은 가장 안정적으로 분류된 반면, 슬픔 - 불안 - 상처, 분노 - 당황, 기쁨 - 중립 감정 쌍에서 상대적으로 높은 혼동이 관찰되었다. 이러한 결과는 얼굴 표정 기반 감정 분류의 한계를 반영하는 동시에, 실제 정서 표현의 연속성이 모델 예측에 반영된 것으로 해석할 수 있다.

4.5 우울 관련 감정 / 비우울 관련 감정 이진 분류 실험 비교

추가로, 감정 클래스를 우울 관련 감정과 비우울 관련 감정으로 단순화한 이진 분류 실험을 수행하였다. 우울 관련 감정 셋으로는 슬픔, 불안, 상처 데이터를 합쳤고, 비우울 관련 감정 셋으로는 기쁨, 분노, 당황, 중립의 데이터를 합쳤다. 그러나 해당 실험에서는 validation accuracy가 37.2%에 그쳤으며, 모델이 대부분의 샘플을 우울 관련 감정으로 예측하는 심각한 편향 현상이 발생하였다.

Ⅴ. 결론

본 논문에서는 임상적 우울 진단 라벨(PHQ-9)에 의존하지 않고, 얼굴 표정에 내재된 감정 표현을 정밀하게 학습하기 위한 감정 특화 CLIP 기반 시각 표현 모델을 제안하고 그 성능을 분석하였다. 이는 실제 연구 환경에서 임상 라벨 확보가 어려운 상황을 고려한 대안적 접근으로, 우울 관련 비언어적 신호 분석을 위한 선행 단계에 해당한다.

실험 결과, 제안한 모델은 정적 얼굴 이미지 기반 zero-shot 감정 분류에

서 56.28%의 정확도를 기록하였으며, 중립과 슬픔 감정에서 비교적 안정적인 분류 성능을 보였다. 반면, 불안·상처·당황 감정은 슬픔 또는 중립으로의 혼동이 빈번하게 발생하였으며, 혼동 행렬 분석을 통해 감정 간 구분이 얼굴 표정만으로는 제한적일 수 있음을 확인하였다. 또한 감정을 우울 관련 감정과 비우울 관련 감정으로 단순화한 이진 분류 실험에서는 심각한 예측 편향이 발생하여, 추상적인 이진 라벨이 CLIP 기반 의미 정렬 학습에 적합하지 않음을 확인하였다.

본 연구의 의의는 감정 분류 성능 자체보다는, 우울과 연관된 감정 표현을 포함하는 시각적 의미 공간을 안정적으로 학습할 수 있는 기반 표현 모델을 구축했다는 점에 있다. 향후 연구에서는 본 모델을 기반으로 멀티모달 영상 데이터에서 얼굴, 음성, 텍스트 정보를 시간 구간 단위로 정렬하고, 이를 활용한 우울 관련 비언어적 신호 분석을 연구할 예정이다.

ACKNOWLEDGMENT

본 연구는2022년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업의 연구결과로 수행되었음.(2022-0-00964)

참 고 문 헌

[1] World Health Organization. ("August 29 "). *Depressive disorder (depression).* Available: https://www.who.int/news-room/fact-sheets/detail/depression.
[2] 감수일. ("May 13, "). *우울과 우울장애.* Available: https://www.mentalhealth.go.kr/portal/disease/diseaseDetail.do?dissId=38.
[3] 조은의 (Eunui Jo) and 김제중 (Jejoong Kim), "우울증의 얼굴표정 정서정보 처리 연구에 대한 통합적 문헌고찰," *스트레스硏究,* vol. 28, *(2),* pp. 41 - 49, 2020. .
[4] H. R. 류경희 / Kyoung and J. O. 오경자 / Kyung, "우울감이 얼굴 표정 정서 인식에 미치는 영향," *감성과학 / Science of Emotion & Sensibility,* vol. 11, *(1),* pp. 11 - 21, 2008. .
[5] S. Li and W. Deng, "Deep Facial Expression Recognition: A Survey," *IEEE Transactions on Affective Computing, Affective Computing, IEEE Transactions on, IEEE Trans.Affective Comput.,* vol. 13, *(3),* pp. 1195 - 1215, 2022. .
[6] X. Zhang, B. Li and G. Qi, "A novel multimodal depression diagnosis approach utilizing a new hybrid fusion method," *Biomedical Signal Processing and Control,* vol. 96, pp. 106552, 2024. Available: https://www.sciencedirect.com/science/article/pii/S1746809424006104. DOI: 10.1016/j.bspc.2024.106552.
[7] Y. Li*et al*, "Predicting depression by using a novel deep learning model and video-audio-text multimodal data," *Front. Psychiatry,* vol. 16, 2025. . DOI: 10.3389/fpsyt.2025.1602650.
[8] A. Radford*et al*, "Learning Transferable Visual Models From Natural Language Supervision," .