비대면 수업에서의 학습자 감정 인식을 위한 딥러닝 기반 다중 모달 시스템

윤혜원, 김혜지*, 박종열*
서울과학기술대학교

yoonhw@seoultech.ac.kr, *kimhyejee923@seoultech.ac.kr, *jongyoul@seoultech.ac.kr

Deep Learning-Based Multimodal System for Emotion Recognition in Learners during Online Classes

Yoon Hyewon, Kim Hyeji*, Park Jongyoul*
Seoul National Univ., of Science and Technology

요 약

코로나 19 팬데믹으로 비대면 수업이 증가하면서 학생들의 감정 상태와 학습 동기를 정확히 파악하기 어려워졌다. 이를 해결하기 위해 인공지능(AI) 기반 감정 인식 기술이 주목받고 있으며, 특히 다중 모달 접근법이 중요하다. 본 논문은 얼굴 표정과 신체 자세를 결합하여 학생들의 감정을 보다 포괄적으로 분석하는 프레임워크를 제안한다. 제안된 프레임워크는 Adaptive Fusion Network 를 통해 얼굴과 신체 정보를 통합하여 학생의 감정 상태를 정확히 분류하고, 이를 바탕으로 실시간 맞춤형 피드백을 제공하는 AI 기반 교육 도우미의 가능성을 탐구한다.

Ⅰ. 서 론

코로나 19 팬데믹으로 인해 비대면 수업이 급증하며 교육 현장에서 새로운 도전 과제가 등장했다. 대면 수업에서는 교수자와 학생 간의 직접적 상호작용을 통해 학생의 감정 상태와 학습 동기를 파악할 수 있었으나, 비대면 수업에서는 이러한 상호작용이 제한되어 학습 효과를 정확히 파악하기 어려워졌다. 이를 해결하기 위해 AI 기반 감정 인식 기술이 주목받고 있으며, 이 기술은 실시간으로 학생의 감정을 모니터링하고 맞춤형 피드백을 제공하는 도구로 활용되고 있다[1]. 감정은 학습 동기와 성과에 중요한 영향을 미치므로, 긍정적 감정은 학습 동기를 높이고, 부정적 감정은 학습을 저해할 수 있다.

하지만, 기존의 감정 인식 연구는 주로 얼굴 표정에 의존해 감정의 복합적 특성을 충분히 반영하지 못한다. 감정은 얼굴 표정뿐만 아니라 시선, 신체 자세, 손동작, 음성 톤 등의 비언어적 신호를 통해 표현되며[2], 이를 효과적으로 해석하려면 다중 모달 접근법이 필요하다. 단일 모달 접근법은 특정 신호에 집중해 중요한 단서를 놓칠 수 있으나, 다중 모달 접근법은 다양한 신호를 통합해 더 포괄적이고 정밀한 감정 인식을 가능하게 한다. 예를 들어, 얼굴 표정만으로는 학생의 복잡한 감정을 완전히 파악하기 어렵지만, 신체 자세와 손동작을 함께 분석하면 더 정확한 이해가 가능하다.

본 논문은 이러한 다중 모달 정보를 활용해 교육 환경에서 학생들의 감정 상태를 분류하는 프레임워크를 제안한다. 얼굴 표정과 신체 자세를 결합해 학생의 감정을 포괄적이고 맥락적으로 이해함으로써, 실제 교육 환경에서 발생하는 미묘한 감정 신호를 정확하게 해석하고, AI 기반의 맞춤형 교육 도우미를 통해 즉각적이고 개인화된 피드백을 제공할 수 있는 기반을 마련하고자 한다.

Ⅱ. 다중 모달 감정 인식 프레임워크

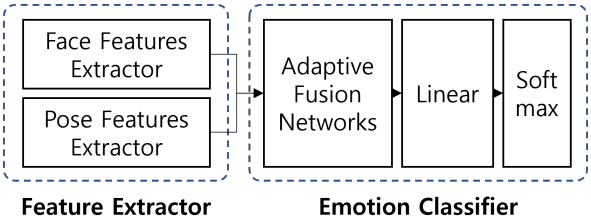

그림 1. 다중 모달 감정 인식 프레임워크

본 논문은 얼굴 특징과 신체 자세 특징을 활용하여 학생들의 심리적 상태를 분류하는 프레임워크를 제안한다. 이 프레임워크는 특징 추출기와 감정 분류기로 구성된다.

특징 추출기는 크게 얼굴 특징 추출기와 신체 자세 특징 추출기로 구성된다. 먼저, 얼굴 특징 추출기는 얼굴 영역의 특징을 분석하기 위한 부분으로, 얼굴 이미지를 탐지하고 이를 잘라낸 후, 얼굴 표정 특징을 추출하는 역할을 한다. 이 과정에서, 얼굴의 랜드마크(눈, 코, 입 등)를 검출하고, 얼굴 이미지 내에서 중요한 부분을 강조하기 위해 cross-attention 메커니즘을 적용한다. 이는 얼굴 표정에서 중요한 세부 사항을 더 잘 포착할 수 있게 하며, 특히 감정 인식을 향상시키는 데 중요한 역할을 한다. 신체 자세 특징 추출기는 얼굴 이외의 신체 정보를 분석하기 위한 부분으로, 상반신 스켈레톤을 GCN(Graph Convolutional Networks)[3]에 입력하여 신체 자세 특징을 추출한다. 이러한 신체 정보를 통해 학생의 심리적 상태를 보다 정확하게 파악할 수 있다. 예를 들어, 학생이 긴장하거나 불안할 때 나타나는 특정한 신체적 움직임이나 자세 변화를 감지할 수 있다.

	긍정		부정	
	활성	비활성	활성	비활성
활동	Enjoyment	Relaxation	Anger	Boredom
결과	Joy	Contentment	Anxiety	Sadness
	Hope	Relief	Shame	Hopelessness
	Pride		Anger	Disappointment
	Gratitude			

표 2. 교육 환경에서의 16 개 감정

감정 분류기는 교육 환경에서 학생들의 다양한 심리적 상태를 정확히 분류하여, 최종적으로 학생 개개인에게 맞춤형 교육 피드백을 제공하는 데 기여한다. 교육 환경에서 학생들의 감정은 3 가지 차원으로 분류된다: 긍정적/부정적 감정(밸런스), 활성화된/비활성화된 감정(활성화 정도), 그리고 감정의 대상(활동 또는 결과). 이에 따라 교육 환경에서 학생들의 감정은 표 1 과 같이 16 가지로 분류된다. 감정 분류기는 Adaptive Fusion Network 와 선형 레이어, 그리고 Softmax 로 구성된다. 먼저, Adaptive Fusion Network 는 얼굴 표정과 신체 자세의 두 가지 주요 모달리티에서 추출된 특징 벡터를 입력으로 받아, 각 모달리티의 중요도를 동적으로 조정하는 메커니즘을 통해 감정 상태를 분류한다. 각 모달리티의 입력 특징 벡터는 각각의 선형 변환 계층을 통해 은닉 차원의 임베딩 벡터로 변환된다. 이후, 각 모달리티에 대해 별도로 학습된 attention 가중치가 이 임베딩 벡터에 적용된다. 이 가중치들은 학습 과정에서 각 모달리티의 상대적 중요도를 반영하여 조정된다. 이렇게 가중치가 반영된 두 임베딩 벡터는 합산되어 하나의 융합된 특징 벡터를 형성한다. 이 융합된 특징 벡터는 두 개의 선형 변환 계층과 비선형 활성화 함수로 구성된 분류 모듈에 입력된다. 첫 번째 선형 변환 계층은 융합된 특징 벡터를 은닉 차원에서 처리하며, 비선형 활성화 함수를 통해 모델의 표현력을 높인다. 두 번째 선형 변환 계층은 최종 출력 차원으로 변환하여 감정 상태를 예측하는 역할을 수행한다. 최종 출력은 Softmax 함수에 의해 각 감정 클래스에 속할 확률 분포로 변환된다. 이 확률 분포는 입력된 데이터가 각 감정 클래스에 속할 확률을 나타내며, 가장 높은 확률을 가지는 클래스가 최종적으로 선택된다. Adaptive Fusion Network 는 다중 모달 정보의 복잡한 상호작용을 효과적으로 처리하도록 고안되었으며, 특히 attention 메커니즘을 통해 모달리티 간의 정보 불균형을 보완하여 통합된 특징 표현을 만들어낸다.

Ⅲ. 실험

본 연구에서는 제안된 다중 모달 감정 인식 프레임워크의 성능을 검증하기 위해 다양한 실험을 수행하였다. 실험에서는 Adaptive Fusion Network 가 각 모달리티 간의 중요도를 동적으로 조정하여 최적의 감정 분류 성능을 발휘하는지 확인한다. 실험에 사용된 데이터셋은 실제 교육 환경을 모방한 시뮬레이션 데이터로 구성되었다. 실험자는 16 가지 감정을 표현하도록 요청받았으며, 이 과정은 비디오로 녹화되었다. 실험 과정은 다음과 같이 진행되었다. 첫 째, 첫째, 수집된 데이터를 전처리하여 모델 학습에 적합한 형태로 변환하였다. 얼굴 표정 특징의 경우, 계산량을 줄이기 위해 각 프레임에서 MTCNN(Multi-task Cascaded Convolutional Networks) 알고리즘[4]을 사용하여 얼굴 영역을 인식하고, 인식된 얼굴 영역을 이미지 파일로 저장하였다. 신체 자세 특징은 RGB 이미지로부터 3D 스켈레톤 좌표를 추출하여 json 파일로 저장하였다. 둘 째, 제안된 프레임워크와 단일 모달 프레임워크를 비교한다. 학습 과정에서는 교차 검증 기법을 사용하여 모델의 성능을 평가하였다. 모델의 성능은 분류 정확도(accuracy)로 측정되었으며, 각 감정 클래스에 대한 분류 정확도와 학습 정확도를 기록하였다.

Ⅳ. 실험 결과 및 결론

실험 결과, 제안된 다중 모달 감정인식 프레임워크는 16 개의 감정 분류 정확도로 72.98%를 기록하였다. 얼굴 표정 단일 모달은 분류 정확도로 71.50%를 기록하였다.

이와 같은 결과는 제안된 다중 모달 감정 인식 프레임워크가 교육 환경에서 학생들의 복합적인 감정 상태를 보다 효과적으로 파악할 수 있음을 시사한다. 특히, Adaptive Fusion Network 는 얼굴 표정과 신체 자세 간의 정보 불균형을 효과적으로 보완하여 감정 상태를 정확하게 분류할 수 있었다. 이러한 프레임워크는 AI 기반의 맞춤형 교육 도우미(AI Copilot) 시스템에 적용되어, 학생들의 학습 과정에서 실시간으로 적절한 피드백을 제공하는 데 기여할 수 있을 것으로 기대된다. 향후 연구에서는 다양한 교육 상황과 감정 상태를 고려하여 프레임워크의 적용 범위를 확장하고, 실제 교육 현장에서의 활용 가능성을 검증할 필요가 있다. 또한, 감정 인식의 정확도를 더욱 향상시키기 위해 시선 추적, 음성 인식 등 추가적인 모달리티를 통합하는 연구가 필요할 것이다.

ACKNOWLEDGMENT

본 논문은 교육부와 한국연구재단의 재원으로 지원을 받아 수행된 첨단분야 혁신융합대학사업의 연구결과입니다. (No.2024-0762, 인공지능 기반의 학습자 몰입도 평가에 관한 연구) & 본 연구는 문화체육관광부 및 한국콘텐츠진흥원의 2024 년도 문화체육관광 연구개발사업으로 수행되었음 (과제명 : 시니어의 콘텐츠 제작 접근성 향상을 위한 생성형 AI 기반 콘텐츠 창·저작 플랫폼 기술 개발, 과제번호 : RS-2024-00340342, 기여율: 30%)

참 고 문 헌

[1] Ngo, Duong, et al. (2024). Facial expression recognition for examining emotional regulation in synchronous online collaborative learning. *International Journal of Artificial Intelligence in Education*, pp. 1-20.

[2] Kim, Seok Min, et al. (2022-06-22). Deep learning-based Korean speech emotion recognition. *Proceedings of Symposium of the Korean Institute of communications and Information Sciences*, Jeju.

[3] Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

[4] Zhang, Kaipeng, et al. (2016). Joint face detection and alignment using multitask cascaded convolutional networks. IEEE signal processing letters, 23(10), 1499-1503.