DBpia

멀티모달 딥러닝을 활용한 감정 분류 연구

임서연[O2] 차수정[2] 최유진[2] 동서연[1]
[1]숙명여자대학교 인공지능공학부
[2]숙명여자대학교 IT공학전공
imseoyeon0218@sookmyung.ac.kr, soojeongcha@sookmyung.ac.kr, cuttie1123@sookmyung.ac.kr, sydong@sookmyung.ac.kr

Multi-modal Deep Learning for Emotion Classification

Seoyeon Lim[O2] Soojeong Cha[2], Eugene Choi[2], Suh-Yeon Dong[1]
[1]Department of Artificial Intelligence Engineering, Sookmyung Women's University
[2]Department of Information Technology Engineering, Sookmyung Women's University

요 약

최근 감성 컴퓨팅(Affective Computing) 분야에서는 멀티모달 데이터셋을 활용하여 감정인식을 수행하고자 하는 연구가 늘어나고 있다. 단일 모달 데이터셋은 다양한 감정 요소를 포착하는 데 한계가 있는 반면에, 멀티모달 데이터셋을 사용하면 더 다양하고 복잡한 감정요소를 파악할 수 있으며, 새로운 모달을 추가하여 활용이 가능하다. 따라서 본 연구에서는 KEMDy19 데이터셋의 음성, 텍스트, 심전도, 피부전도도, 손목 피부온도 데이터를 사용해 7가지 감정(기쁨, 놀람, 분노, 중립, 혐오, 공포, 슬픔)을 예측하는 멀티모달 딥러닝 기반 감정 분류를 수행한다. 또한 멀티모달 데이터셋에서 다양한 모달의 조합으로 감정 분류를 수행한 결과를 제시하여 멀티모달 감정 분류에서의 생체 신호의 유용성을 확인하고자 한다. 연구의 결과는 생체신호를 발화 데이터와 함께 분류에 사용했을 경우 발화 데이터만 분류에 사용한 경우보다 높은 accuracy와 f1 score를 보였다. 본 연구의 결과는 감정 분야에서 딥러닝 기반 멀티 모달 데이터셋의 활용과 사람의 감정을 보다 정확하게 분류하기 위해 생체 신호의 사용을 제안하는 근거를 제시한다.

1. 서 론

감정 분류는 사람들의 감정 상태를 파악하고 이를 기계적으로 분류하는 분야로, 자연어 처리, 컴퓨터 비전, 음성 인식 등 다양한 분야에서 활용되고 있다. 최근에는 기존의 텍스트나 음성 데이터 뿐만 아니라 다양한 모달리티(modality)를 가진 데이터셋을 사용하여 감정 분류를 수행하는 연구들이 증가하고 있다[1,2]. 생체 신호란 인간의 생리적인 반응을 나타내는 신호로서, 심전도(ECG; Electrocardiogram), 피부전도도(EDA; Electrodermal Activity), 손목 피부온도 등이 있다. 이러한 생체 신호는 인간의 감정 상태를 직접적으로 반영하여 보다 정확하고 신뢰성 높은 감정 분류 결과를 얻을 수 있다. 예를 들어, 피부 전도도는 스트레스와 같은 감정 상태와 밀접한 관련이 있기 때문에 이를 이용한 감정 분류 연구가 이루어져 왔다[3]. 또한, 생체 신호는 비언어적인 정보를 제공하므로, 텍스트나 음성 데이터 등 언어적인 정보만을 이용할 때와는 다른 정보를 제공할 수 있다. 더하여, 생체 신호는 비침습적이고 비통증적인 측정이 가능하여 피험자의 부담이 적어 보다 자연스러운 상황에서 감정 분류를 수행할 수 있으며, 이를 활용하여 의료나 심리학 분야의 연구에서도 활용되고 있다[4,5]. 하지만, 생체 신호만을 이용한 감정 분류는 비슷한 생체 신호 패턴이 여러 가지 감정 상태에[1]서 나타날 수 있다는 한계점이 존재한다.

따라서 최근에는 다양한 모달을 이용한 감정 분류 연구가 진행되고 있다. 텍스트와 오디오 데이터는 감정 상태를 표현하는 언어적인 정보를 제공하며, 이러한 정보를 이용하여 감정 분류를 수행하는 연구가 이루어져 왔다[6,7,8]. 또한, 멀티 모달 데이터셋에서 다양한 모달을 조합하여 감정 분류를 수행하는 연구도 이루어져 왔다. 예를 들어, 얼굴 표정, 음성 조언, 자연언 표현 등 다양한 모달리티를 모두 활용하여 분석하면 생체 신호만을 사용할 때보다 향상된 감정 분류 결과를 얻을 수 있다[9]. 본 연구에서는 감정 분류에 적합한 모달의 조합을 제시하는 것을 목적으로 한다. 이를 위해 생체 신호, 텍스트, 오디오 멀티 모달 데이터셋인 KEYMDy19에서 다양한 모달을 조합하여 감정 분류를 수행한다. 오디오와 텍스트, 복합적인 생체 신호를 분석하기 위하여 딥러닝을 적용하고, 최종적으로 다중 모달 생체신호를 통한 딥러닝 기반 감정 분류기를 설계한다. 본 연구에서 제안한 감정 분류기를 활용하여 도출한 결론은 멀티모달 생체신호의 유용성을 입증하는 근거로 쓰일 수 있을 것이다.

2. 방 법

2.1 데이터셋

한국어 멀티모달 감정 데이터셋(KEMDy19; Korean Emotional Multi-modal Dataset in 2019)[10]은 발화 음성, 발화의 문맥적 의미, 생체 신호(ECG, EDA, 손목 피부온

* This work was supported by the National Research Foundation of Korea (NRF) grand funded by the Korea government (MSIT) (No. NRF-2021R1F1A1052389)

www.dbpia.co.kr

도) 등을 포함하고 있는 멀티모달 감정 데이터셋이다. 40명의 한국인 성우를 대상으로 총 20개의 세션에서 10개의 감정 상황극을 연기하였고, 이 과정에서 발화 음성, 발화 텍스트, 발화자의 생체 신호 데이터를 수집하였다. 감정 레이블은 외부 관찰자 10명이 녹화된 감정 상황극 영상을 시청한 후, 발화 세그먼트 별로 7가지 카테고리 감정 레이블과 각성도, 긍/부정도를 평가하여 가장 많이 선택된 레이블로 결정되었다. 데이터셋의 80%는 Training에 20%는 Test에 사용했다.

2.2 특징 추출

심전도의 시계열 데이터에서 R 피크 정보로부터 계산되는 심박변이도(HRV; Heart Rate Variability)의 평균(mean HRV), 표준편차(SDNN; Standard Deviation of NN Intervals), RMSSD(Square Root of the Mean Squared Difference of Successive NNs), NN50(Number of Pairs of Successive NNs that Differ by More Than 50ms)를 특징으로 추출하였다. 피부 전도도 신호는 푸리에 변환을 거쳐 주파수 스펙트럼으로 변환시키고, 0.05-5Hz의 주파수 범위를 관심 범위로 설정하여 이 범위 내의 주파수 구성 요소를 특징으로 추출하였다. 손목 피부온도는 시간적인 특성을 분석하여, 시간 영역 특징인 평균값, 분산, 기울기, y-절편을 분류 특징으로 추출하였다.

발화 텍스트를 감정 분류에 사용하기 위해서는 텍스트 데이터를 토큰화하고, 특수문자나 불필요한 문자열을 제거하며, 텍스트를 정규화하는 전처리 과정이 필요하다. 본 연구에서는 KoBERT 모델의 토크나이저를 사용하여 한글 문장을 단어나 문장으로 나누는 토큰화 작업을 수행했고, 각 토큰을 숫자로 매핑하고 텐서로 변환하여 특징으로 추출했다. 각 토큰의 패딩 여부 정보를 포함한 어텐션 마스크(attention mask) 또한 추출하여 모델의 입력으로 사용했다. 마지막으로, 발화 음성은 PyTorch의 음성 처리 패키지인 torchaudio를 사용하여 해당 음성 파일을 텐서 형태로 변환하여 모델의 입력으로 사용했다.

2.3 분류기

본 연구에서는 생체신호, 발화 음성과 발화 텍스트 정보를 결합하는 방법으로 late fusion 방식을 사용했다. Late fusion은 여러 모달리티부터 추출된 특징을 개별로 처리한 후 결합하는 방법으로, 모델이 모든 모달리티의 정보를 동시에 처리할 수 있는 복잡한 모델링을 필요로 하지 않기 때문에, 계산 비용이 낮아지고 모델 구성이 단순해지는 장점을 가지고 있다. 본문에서는 생체신호와 오디오, 텍스트 개별 특징을 처리하기 위해 사용한 모델에 대해 설명하고, late fusion을 거친 최종 모델을 제시한다.

2.3.1 생체신호

생체 신호를 사용한 감정 분류 모델의 아키텍처는 완전 신경망(Fully Connected Neural Network)으로 구성되었다. 완전 신경망은 딥러닝에서 가장 기본적인 형태의 인공 신경망 모델 중 하나로, 모든 뉴런이 이전 층의 모든 뉴런과 연결되어 있는 구조를 가지고 있다. 모델은 3개의 fully connected layer로 이루어져 있으며 활성화 함수로는 ReLu 함수를 사용했다. 출력층에서 Softmax 함수가 활성화 함수로 사용되었으며, batch size는 4, 학습률은 0.001, epoch는 50번으로 설정하여 학습을 진행하였다.

2.3.2 발화 음성

발화 음성을 사용한 감정 분류 모델의 아키텍처는 합성곱 신경망(CNN; Convolutional Neural Network)과 완전 연결 신경망(Fully Connected Neural Network)로 구성되었다. 모델의 입력은 1D 합성곱 레이어를 통해 처리된다. 이후, 출력된 특징 맵(feature map)은 1차원 벡터로 평탄화되어 완전 연결 신경망으로 전달된다. 출력층에 활성함수를 추가하는 대신, 모델 훈련에서 CrossEntropyLoss 함수가 사용되어 모델의 출력값에 대한 소프트맥스 함수와 네거티브 로그 가능도(negative log likelihood)를 결합해준다. batch size는 32, 학습률은 0.001, epoch는 50번으로 설정하여 학습을 진행하였다.

2.3.3 발화 텍스트

발화 텍스를 사용한 감정 분류 모델로 KoBERT [11]가 사용되었다. KoBERT는 한국어 자연어 처리를 위해 사전 학습된 언어 모델이며 구글의 BERT(Bidirectional Encoder Representations from Transformsers) 모델 아키텍처를 기반으로 한다. 모델은 입력 텍스트의 양쪽 방향에서 정보를 인코딩하는 양방향 인코더로 구성되었고, 각 인코더는 12개의 레이어로 구성된 트랜스포터 아키텍처를 사용한다. batch size는 32, 학습률은 0.001, epoch는 50번으로 설정하여 학습을 진행하였다.

2.3.4 멀티 모달

위에서 설명한 3종류의 단일모달 감정 분류 모델들은 입력 데이터를 독립적으로 처리하고 각 모델이 생성한 출력을 조합하여 최종 예측을 만들어내는 late fusion 방법을 사용해 감정 분류를 수행한다. 이 방법은 다양한 종류의 모델 또는 다른 유형의 입력을 사용하여 결과를 결합할 수 있을만큼 확장성이 높다. 본 연구에서는 사전 학습된 3개의 모델을 불러온 다음, 각 모델의 출력값을 가져와 late fusion을 적용했다.

3. 결 과

감정 분류에서 생체 신호의 효과를 확인하기 위하여 생체 신호인 심전도, 피부 전도도, 손목 피부온도와 함께 발화 음성, 텍스트를 조합하여 멀티모달 데이터셋으로 감정 분류를 진행하였다. 성능 평가 지표로는 Accuracy와 F1-score를 사용하였다. 발화 음성인 Audio와 발화 텍스트인 Text를 단일 모달로 사용해 분류

www.dbpia.co.kr

예측을 진행한 결과, 각각 61%, 41%의 accuracy를 보였으며 f1 score는 0.39와 0.26을 보였다. Audio와 Text를 조합한 멀티모달 데이터셋으로 감정 분류를 진행했을 때 3가지 생체신호를 조합하여 감정 분류를 진행했을 경우, 52%의 accuracy와 0.31의 f1 score를 보였다. KEDMy19 데이터셋에서 제공하는 다섯 모달을 모두 분류에 사용한 경우, 70%의 accuracy와 0.49의 f1-score로 가장 높은 분류 성능을 보여줬다.

Setting	Accuracy	F1-Score
Audio	0.61	0.39
Text	0.41	0.26
Bio Signals	0.52	0.31
Audio+Text	0.63	0.42
Audio+Text +Bio Signals	**0.70**	**0.49**

표 1 생체 신호, 발화 음성과 발화 텍스트를 사용한 분류 성능

4. 결 론

본 연구는 감정 분류에서의 생체 신호의 유용성을 확인하기 위해 KEMDy19 데이터셋으로부터 심전도, 피부 전도도, 손목 온도, 발화 음성 및 텍스트 정보를 받아 멀티모달 딥러닝 기반 감정 분류를 수행했다. 그 결과, 생체 신호와 발화 데이터를 함께 사용한 경우에 70%의 accuracy로 가장 높은 성능을 보여줬으며, 이는 발화 데이터를 사용한 결과와 비교하여 생체 데이터의 포함 여부가 분류 성능에 영향을 미침을 의미한다. 따라서 본 논문에서는 생체 신호가 감정 분류에 효과적인 모달임을 입증하였으며, 감정 분류가 필요한 다양한 분야에서의 생체 신호 사용을 제안한다.

5. 참고문헌

[1] Catherine M, Dariusz M, Krzysztof T, Piotr P, Lamine B, Corinne A, Katarzyna W, “ Survey on ai-based multimodal methods for emotion detection”, Springer, High-Performance Modelling and Simulation for Big Data Applications. p307-324, 2019

[2] Koelstra S, Mühl C, Soleymani M, Lee JS, Yazdani A, Ebrahimi T, “ Deap: A database for emotion analysis; using physiological signals, IEEE transactions on affective computing”, IEEE transactions on affective computing, Vol. 3, No. 1, p18-31, 2011

[3] Nasoz F, Alvarez, Lisetti, C, “ Emotion recognition from physiological signals using wireless sensors for presence technologies”, Cogn Tech Work Vol. 6, p4-14, 2004

[4] Anneketh V, Jyotika P, “ An automated psychometric analyzer based on sentiment analysis and emotion recognition for healthcare”, Procedia computer science, Vol. 132, p1184-1191, 2018

[5] Azam N, Ahmad T, Haq N, “ Automatic emotion recognition in healthcare data using supervised machine learning”, PeerJ Computer Science, Vol. 7, e751, 2021

[6] Li W, Xu H, “ Text-based emotion classification using emotion cause extraction”, Expert Systems with Applications, Vol. 41, Issue 4, Part 2, p1742-1749, 2014

[7] Bhaskar J, Sruthi K, Nedungadi P, “ Hybrid approach for emotion classification of audio conversation based on text and speech mining”, Procedia Computer Science, Vol. 46, p635-643, 2015

[8] Noh KJ, Jeong CY, Lim J, Chung S, Kim G, Lim JM, Jeong H, “ Multi-Path and Group-Loss-Based Network for Speech Emotion Recognition in Multi-Domain Datasets”, Sensors, Vol. 21, No.5, 1579, 2021

[9] Cimtay Y, Ekmekcioglu E, Caglar-Ozhan S, “ Cross-subject multimodal emotion recognition based on hybrid fusion”, IEEE Access, Vol. 8, p168865-168878, 2020

[10] K. J. Noh and H. Jeong, “ KEMDy19,” https://nanum.etri.re.kr/share/kjnoh/KEMDy19?lang=ko_KR

[11] Devlin J, Chang MW, Lee K, Toutanova K, “ Bert: Pre-training of deep bi-directional transformers for language understanding”, arXiv, 2018

www.dbpia.co.kr