각성도 및 긍/부정도의 싱글모달 사전 학습 예측 모델 기반 멀티모달 감정인식 모델

홍지우○, 김예찬, 윤지영, 채소연, 한지원
성균관대학교 통계학과
jiwoo1000@g.skku.edu, dmalbarn@g.skku.edu, jyggang@g.skku.edu,
cothdus28@g.skku.edu, jiwon.h@g.skku.edu

Multimodal Emotional Recognition Model based on Singlemodal Pretrained Prediction Model of Valence and Arousal

Jiwoo Hong○, Yechan Kim, Jiyoung Yoon, Soyeon Chae, Jiwon Han
Department of Statistics, SungKyunKwan University

요 약

본 연구에서는 발화자의 음성 및 텍스트, 그리고 각성도(Arousal), 긍/부정도(Valence)를 반영하여 감정을 분류하는 멀티모달 모델을 제시한다. 발화자의 음성에서 각성도를, 텍스트에서 긍/부정도를 반영할 수 있도록 사전 학습 모델을 기반으로 특징 추출 과정을 구성하여 기존의 멀티모달 모델들과 차별점을 갖는 모델을 구축하였다. 그 결과, 정확도 89.82%, F1 Score 86.93%의 감정 분류 성능을 확인하였다.

1. 서 론

비언어적(non-verbal) 요소는 언어적(verbal) 요소 못지않게 발화의 내용과 감정의 전달에 관여한다[1]. 각성도와 긍/부정도는 비언어적 요소의 대표적인 사례이며, 발화자의 감정과 유의한 관계가 존재함이 연구된 바 있다[2]. 따라서 본 연구에서는 언어적 요소인 음성과 텍스트를 기반으로 감정을 인식하는 멀티모달 모델에 비언어적 요소인 각성도와 긍/부정도를 반영함으로써 정확도를 높였다.

본 연구에서 제안하는 모델은 트랜스포머(Transformer) 모델[3]의 인코더(Encoder)를 응용한 멀티모달 딥러닝 모델로, 3개의 블록으로 이루어진다. 첫 번째인 음성 특징 추출 블록은 사전 학습된 음성 기반 각성도 예측 모델의 파라미터를 기반으로 구성한다. 두 번째인 텍스트 특징 추출 블록은 사전 학습된 텍스트 기반 긍/부정도 예측 모델의 파라미터를 기반으로 구성한다. 마지막으로 감정인식 블록은 혼합된 특징들에서 특성을 추출하고, 감정 분류를 진행하는 레이어로 구성된다.

모델의 우수성을 입증하기 위해 데이터 불균형의 극복 가능성과 예측 성능을 주된 평가 요소로 설정했다. F1 Score와 정확도로 분류성능을 평가하였으며, 그 결과 정확도 기준 89.82%, F1 Score 기준 86.93%의 성능을 확인하였다.

2. 데이터 구조 파악/선택

2.1 데이터 구조 파악

사용한 데이터는 '한국어 멀티모달 감정 데이터셋 2020 (KEMDy20)'[1]으로 80명의 일반인 대상자들의 6개 주제에 대한 자유 발화 과정에서 발화 음성, 텍스트, 생체신호를 수집한 데이터이다. 10명의 평가자는 세그먼트별로 감정, 각성도, 긍/부정도를 평가하였고 감정은 최다 선택된 분류가, 각성도와 긍/부정도는 평균값이 레이블로 설정되었다[4].

감정 레이블은 7가지(기쁨, 놀람, 분노, 중립, 혐오, 공포, 슬픔)로 구성되어 있다. '혐오', '공포' 레이블의 경우, 데이터 수가 100개 미만으로, 학습에 충분하지 않다고 판단되어 나머지 5가지(기쁨, 놀람, 분노, 중립, 슬픔) 감정을 선택하여 사용하였다.

2.2 멀티모달 데이터 선택

생체신호의 변화가 발화자의 감정 분류에 있어 유의한 영향을 주는지 파악하기 위해 생체신호 데이터 중 기록의 주기성과 데이터의 완결성을 유일하게 모두 만족하는 피부 온도 데이터에 대해서 시계열 군집화를 진행하였다.

발화 음성의 특성상 데이터의 길이가 모두 상이하므로 DTW (Dynamic Time Warping) 시계열 군집화를 적용하였다. 군집 수에 따른 Davies-Bouldin Index(DBI)와 Silhouette Index를 비교분석한 결과와 라벨의 가짓수를 종합하여 최적의 클러스터 개수를 5개로 판단하고 이에 대해 그림 1과 같이 Silhouette Analysis를 진행하였다.

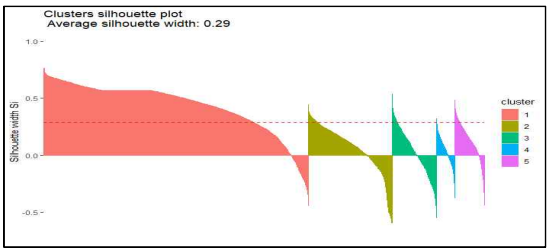

그림 1 Silhouette Analysis 결과

분석 결과, 개별 실루엣 계수(Silhouette Coefficient)에 음수인 값이 다수 존재하고 전체 데이터의 실루엣 계수 평균값이

1 https://nanum.etri.re.kr/share/kjnoh/KEMDy20?lang=ko_KR

0.5 미만의 값을 가지므로 피부 온도는 감정과 유의미한 관계가 없다고 판단하였다. 따라서 텍스트와 음성 데이터만을 사용하여 감정인식 모델 연구를 수행하였다.

3. 데이터 전처리

학습 데이터의 라벨 분포는 그림 2와 같다. 왼쪽부터 중립(neutral), 슬픔(sad), 기쁨(happy), 놀람(surprise), 분노(angry)에 해당한다. 데이터 불균형을 해소하기 위해 슬픔, 놀람, 분노 라벨에 대해 각각 250%까지 데이터를 증강하였다.

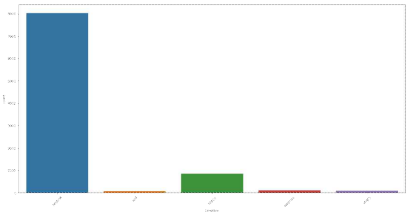

그림 2 학습 데이터 라벨별 분포

3.1 텍스트 데이터 전처리

언어적 표현 이외의 태깅은 모두 제거한 후 KoBERT의 토크나이저로 문장을 토큰화하였다[5]. 이후 토큰화된 문장의 최대 길이를 64로 지정하고 zero-padding을 적용하였다.

3.2 텍스트 데이터 증강

토큰화 단계 이전 텍스트 데이터에 대하여 일정 비율의 단어들을 무작위로 제거(Random Deletion, RD) 및 단어의 순서를 변경(Random Swap, RS)하여 데이터를 증강했다[6].

3.3 음성 데이터 전처리

음성파일을 Mel Spectrogram으로 처리하는 과정을 진행하였다. 음성의 변화를 강조하기 위해 Mel Spectrogram에 대한 1차 차분 값을 함께 활용하였다.

3.4 음성 데이터 증강

Mel Spectrogram 변환 후 일정 주파수 영역을 가리는 Frequency Masking 기법을 적용하였다[7]. 본 연구는 일정한 값으로 특정 영역을 가리는 일반적인 Frequency Masking 대신 랜덤 노이즈를 추가하는 Random Frequency Masking을 그림 3과 같이 도입하여 일반화 성능 향상을 시도하였다.

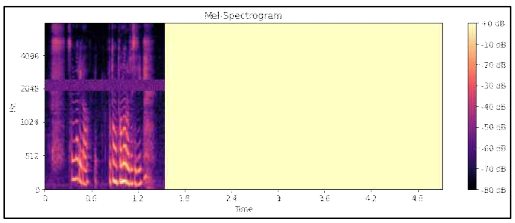

그림 3 음성 데이터 전처리 결과

4. 모델 구축

음성과 텍스트를 사용하는 멀티모달 감정인식 모델에 발화자의 각성도와 긍/부정도를 반영하기 위해 각성도와 긍/부정도 예측 모델을 사전 학습한 후 예측 레이어를 제거하여 각각 음성과 텍스트 특징 추출 블록으로 활용하였다.

따라서, 음성 특징 추출 블록은 음성 임베딩 레이어, 음성 인코더로 구성되며, 텍스트 특징 추출 블록은 텍스트 임베딩 레이어, 텍스트 인코더로 구성된다. 각 예측 모델에서 각성도 예측 레이어와 긍/부정도 예측 레이어를 제거한 것이다. 그리고 감정인식 블록의 경우 혼합 인코더, 감정 예측 레이어로 구성된다.

4.1 음성 특징 추출 블록

음성 임베딩 레이어는 4개의 합성곱(Convolution) 계층으로 구성된다. 각 합성곱 계층에서 2차원 필터의 너비를 1로 설정함으로써 주파수 대역 간 관계를 중심으로 특징을 추출하였다. 그 결과 데이터의 시간 정보는 보존하는 동시에 주파수 대역의 정보를 128차원에서 768차원으로 세분화할 수 있었다.

음성 인코더는 7개의 멀티 헤드 어텐션 레이어로 구성되어 있다. 멀티 헤드 어텐션 레이어는 음성 데이터의 zero-padding 구간에 대해 마스킹(Masking)을 적용한 후 학습되었다.

음성 인코더에서 데이터의 시간 정보를 보존하였으므로 각성도 예측 레이어에 GRU(Gated Recurrent Units)를 사용하였다.

사전 학습한 각성도 예측 모델에서 멀티 헤드 어텐션 레이어 반복 횟수에 따른 각성도 예측 결과는 그림 4와 같다. 반복 횟수가 7회일 때 RMSE가 가장 작게 나타남을 확인할 수 있다.

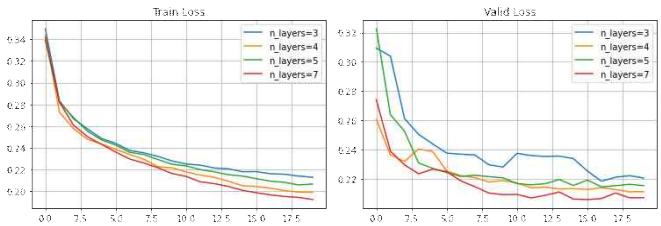

그림 4 음성 멀티 헤드 어텐션 반복수별 RMSE (학습;검증)

4.2 텍스트 특징 추출 블록

텍스트 임베딩 레이어에는 학습 데이터 및 단어 집합의 크기가 작다는 점에서 사전 학습된 KoBERT 모델[5]을 사용하였다.

텍스트 인코더는 5개의 멀티 헤드 어텐션 레이어로 구성되어 있다. 멀티 헤드 어텐션 레이어는 음성 특징 추출 블록과 동일하게 zero-padding 구간에 대해 마스킹을 적용한 후 학습되었다.

텍스트 인코더에서 데이터의 위치 정보를 보존하였으므로, 긍/부정도 예측 레이어에 GRU를 사용하였다.

사전 학습한 긍/부정도 예측 모델에서 멀티 헤드 어텐션 레이어 반복 횟수에 따른 긍/부정도 예측 결과는 그림 5와 같다. 5회 반복했을 때 RMSE가 가장 작게 나타남을 확인할 수 있다.

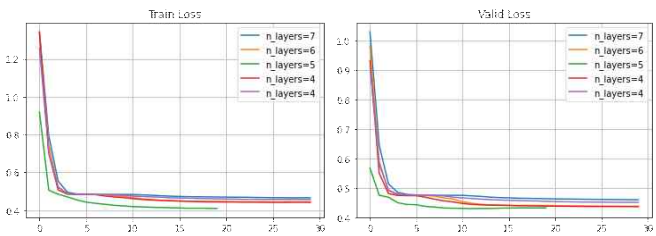

그림 5 텍스트 멀티 헤드 어텐션 반복수별 RMSE (학습;검증)

4.3 감정인식 블록

혼합 인코더는 4개의 멀티 헤드 어텐션 레이어로 구성된다. 혼합 인코더는 음성 인코더, 텍스트 인코더의 출력을 음성, 텍스트 순서로 연결하여 입력으로 사용한다.

멀티 헤드 어텐션 레이어는 병합된 입력에 기존 음성, 텍스트 데이터의 zero-padding 구간을 모두 마스킹한 후 학습되었다.

음성 인코더, 텍스트 인코더, 그리고 혼합 인코더 모두 데이터의 시간•위치 정보를 보존하였으므로, 감정 예측 레이어에 LSTM(Long Short-Term Memory)을 사용하였다.

5. 성능 평가

5.1 F1 Score 비교

총 3가지 (음성 특징 추출 블록, 텍스트 특징 추출 블록) 조합에 대한 모델의 예측 성능을 평가하였다. 모델의 예측 성능은 사전에 분리한 전체 데이터의 20% 데이터로 평가하였다. 표 1은 각 블록의 멀티 헤드 어텐션 레이어 수(n_layers)의 변화에 따른 F1 Score를 비교한 표이다. 순위는 그림 4와 그림 5에서 RMSE가 작게 나타난 n_layers의 순서를 의미한다.

표 1 각 블록의 멀티 헤드 어텐션 반복 횟수 조합별 F1 Score

No.	음성 블록 (n_layers / 순위)	텍스트 블록 (n_layers / 순위)	F1 Score
1	6 / 2nd	6 / 2nd	0.8754
2	7 / 1st	6 / 2nd	0.8760
3	**7 / 1st**	**5 / 1st**	**0.8693**

5.2 Confusion Matrix 비교

1번과 2번 조합의 Confusion Matrix는 그림 6과 같으며, 각 블록의 사전 학습에서 가장 뛰어난 성능을 보인 n_layers 조합으로 학습한 3번 조합의 Confusion Matrix는 그림 7과 같다.

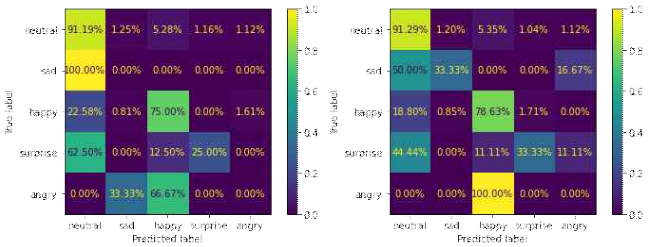

그림 6 멀티모달 모델 Confusion Matrix(1번;2번)

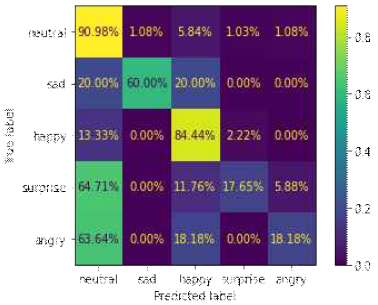

그림 7 멀티모달 모델 Confusion Matrix (3번)

데이터가 적은 슬픔(sad), 놀람(surprise), 분노(angry)에 대한 예측 결과를 비교함으로써 3번 조합의 F1 Score가 소폭 부족하지만 데이터 불균형에 상대적으로 강건함을 확인할 수 있다.

1번과 2번 조합은 모두 분노(angry) 라벨에 대한 예측을 전혀 하지 못했다. 1번 조합의 경우 슬픔(sad)에 대한 예측 또한 하지 못했음을 확인할 수 있다. 반면, 3번 조합은 데이터가 적은 3가지 라벨 모두에 대해 예측 성공 사례가 존재하며, 특히 슬픔(sad)에 대해서는 매우 높은 예측 정확도를 보였다.

6. 결론 및 향후 연구

본 연구는 각성도와 긍/부정도 예측 성능이 뛰어난 모델로 음성과 텍스트의 특징 추출 블록을 구성한다면 데이터 불균형에 강건한 동시에 예측 정확도가 높은 멀티모달 감정인식 모델을 구축할 수 있음을 확인하였다. 따라서, 음성의 경우 각성도를, 텍스트의 경우 긍/부정도를 반영하여 특징을 추출했을 때 더욱 정교한 분류가 가능하다고 판단할 수 있다.

추후 이를 우울증 환자의 감정 연구에 활용한다면 개인정보 문제로 상담 데이터 확보가 어렵다는 점, 우울증 환자의 감정 분포가 상당히 불균형할 것이라는 점을 극복하기 위한 좋은 수단이 될 것이다. 따라서, 우울증 진단 및 발화 분석 연구로 확장한다면 유의미한 결과를 기대할 수 있을 것이다.

참 고 문 헌

[1] 윤애선; 권혁철, “감정 온톨로지의 구축을 위한 구성요소 분석”, 인지과학, 21, 1, 157-175, 2010.

[2] Silke Anders, et al., “Brain Activity Underlying Emotional Valence and Arousal: A Response-Related fMRI Study”, Hum Brain Mapp, 23(4), 200-209, 2004.

[3] Ashish Vaswani, et al., “Attention is all you need”, Proc. of Advances in neural infor-mation processing systems, 5998-6008, 2017.

[4] Noh, K.J., Jeong, C.Y, Lim, J., Chung, S., Kim, G., Lim, J.M., Jeong, H., “Multi-Path and Group-Loss-Based Network for Speech Emotion Recognition in Multi-Domain Datasets”, Sensors, 21(5), 1579, 2021.

[5] SKTBrain (2019). KoBERT, Available: https://github.com/SKTBrain/KoBERT

[6] Jason Wai; Kai Zou, “EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks”, EMNLP-IJCNLP, 2019.

[7] Daniel S. Park; William Chan; Yu Zhang; Chung-Cheng Chiu; Barret Zoph; Ekin D. Cubuk; Quoc V. Le, “SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition”, 2016 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP). IEEE, 2016