MATE : 감정 분석을 위한 오디오-텍스트 혼합 모델

홍성래[O1] 김태미[2] 이솔[1] 김종우[3] 이문용*[1]
[1]카이스트 지식서비스공학대학원
[2]카이스트 산업 및 시스템공학
[3]서울과학기술대학교 산업정보시스템전공
{sun.hong, taemi_kim, leesol4553}@kaist.ac.kr, kjw12316@seoultech.ac.kr, munyi@kaist.ac.kr

MATE : the Multimodal model using Audio and Text for Emotion recognition

SungRae Hong[O1] TaeMi Kim[2] Sol Lee[1] JongWoo Kim[3] MunYong Lee*[1]
[1]KAIST Graduate School of Knowledge Service Engineering
[2]KAIST, Industrial and Systems Engineering
[3]Seoul National University Of Science And Technology, Industrial and Information Systems Engineering

요 약

최근 UX 분야에서 의사소통의 기본 요소인 감정 정보의 중요성이 부각됨에 따라, 딥러닝을 기반으로 감정 인식을 자동화하는 시도가 활발하게 이루어지고 있다. 그러나, 단일 신호만을 고려하는 전통적인 방법은 특정 정보에 의존적인 결과를 생성한다는 한계가 있다. 따라서, 본 논문은 맥락(Context)정보와 멀티모달(Multimodal) 데이터를 기반으로 모델을 제안하였으며, 제안된 모델은 오디오와 텍스트 정보를 동시에 활용하여 감정 정보를 예측한다. 우리는 논멀티모달(Non-Multimodal)과 멀티모달의 성능을 비교 평가하고, 파라미터 탐색을 통해 적합한 성능을 보이는 모델을 선택하고 검증하였다. 실험 결과, 단일 신호에 의존적인 논멀티모달 모델보다 제안된 모델이 더 높은 성능을 보였다.

1. 서 론

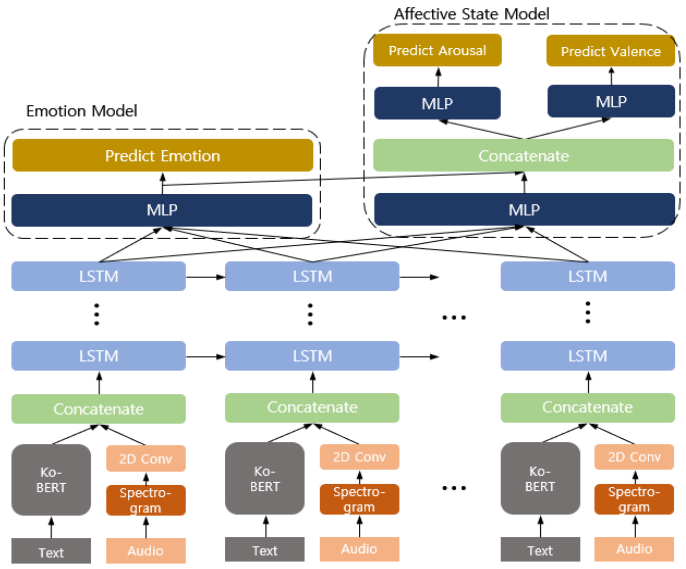
감정 정보는 의사소통의 기본 요소이며, 사람 간 상호작용에 중요한 정보이다. 지난 수년간 사용자 경험연구(User Experience: UX) 분야에서 감정 인식 자동화를 위한 다양한 방법론이 제안되었다. 하지만, 개인은 각기 다른 방식으로 감정을 표현하며, 감정은 시간에 따른 변화가 존재하기에 감정 인식 자동화는 여전히 도전적인 과제로 여겨지고 있다[1]. 이를 극복하기 위해, 최근 인공지능의 발전과 더불어 인간의 감정을 자동으로 인지하는 모델에 대한 연구가 활발하게 이루어지고 있다. 오디오 신호에서 추출된 특징 정보를 Deep Neural Network(DNN)에 학습시켜 음성으로부터 감정을 인식하는 연구[2]와 텍스트 기반의 정보로부터 감정 정보를 인식하는 방법[3] 등이 대표적이다. 하지만, 이러한 단일 신호만을 활용하는 감정 인식 방법은 감정 표현의 복잡성과 시간적 특징을 고려하지 못한다는 한계점을 가지고 있다[4].

기존 단일 신호 방법론들은 단일 데이터 자체의 정보에 의존적이기 때문에 명백한 묘사가 부족한 경우, 주어진 정보에서 정확한 감정을 분석하기 어려웠다. 따라서, 이러한 문제점을 극복하기 위해, 2개 이상의 신호를 동시에 이용하는 Multimodal Emotion Recognition (ER) 연구가 활발하게 진행되고 있다. [5]은 Multi-path 방식을 통해 생성된 특징 벡터를 연결하여 모델 학습에 이용하였다. 그러나 이러한 Multi-path 방식은 텍스트와 오디오의 특징을 별도로 추출함으로써 두 데이터 간의 상호 작용을 반영하기 어렵다는 한계점이 존재한다. 다시 말해, Multi-path 방식은 텍스트, 오디오 각각에 대한 상호 간 보조 정보를 학습하기 어렵다는 것을 의미한다. 따라서, 본 연구는 특징 학습 전 텍스트 임베딩(Embedding)과 오디오 임베딩을 병합 후 LSTM Layer를 통해 하나의 특징 벡터를 생성하는 Single-Path 방식의 감정 분류 모델을 제안한다.

본 논문은 다양한 형태의 멀티모달(Multimodal) 신호를 감정 분석에 적용하기 위해서, 오디오-텍스트 혼합 정보 기반의 감정 분석 모델 "MATE: the Multimodal model using Audio and Text for Emotion recognition"을 제안하며 KEMDy19[5] 데이터 세트로 재현 가능한 소스 코드를 제공한다.

2. 연구 방법

본 연구는 맥락(Context)정보와 멀티모달 데이터를 기반으로 모델을 제안하고자 하며, 제안된 모델은 오디오와 텍스트 정보를 동시에 활용하여 감정 정보를 예측한다. <그림 1>은 본 연구에서 제안하는 오디오-텍스트 멀티 혼합 모델의 구조를 보여준다. 제안된 모델은 오디오와 텍스트 정보의 동시 인식을 위해서, 텍스트 정보와 오디오 정보를 임베딩한 후, 병합하여 LSTM Layer에 투입하도록 구조를 생성했다. 해당 구조를 통해, 동시간대에 발생한 대화의 텍스트와 오디오 정보를 복합적으로 모델이 고려하도록 구성하였다.



<그림 1> 오디오-텍스트 멀티 혼합 모델의 아키텍처

*) 교신 저자

2.1 데이터 전처리 및 입력 데이터

KEMDy19 데이터는 20개의 Session으로 구성되며, 각 Session은 10개 내외의 스크립트(Script)로 이루어진다. 또한 각 사건에 대해서는 다수의 발화가 시간순으로 포함되어있다. 우리는 해당 발화를 8개로 하는 연속된 묶음으로 변경하여, 이를 하나의 대화(Dialogue)로 정의하였다.

데이터의 라벨 종류는 시간 순서에 따른 발화의 텍스트 및 오디오 정보, Speaker와 Listener의 감정(Emotion), 각성도(Arousal, 1~5), 긍/부정도(Valence, 1~5)가 존재한다. 감정은 'happy', 'sad', 'surprised', 'angry', 'neutral', 'disgust', 'fear' 7가지 Class로 분류되어 있으며, 대다수의 대화가 'neutral' Class로 불균형한 라벨을 가지고 있다.

또한, Emotion Labeling의 경우, 평가자의 의견이 엇갈리는 경우들이 다수 존재했다. 우리는 사람의 감정이 하나로 정의하기 어려운 복잡함을 가진다는 가정하에 Hard Labeling 방식이 아닌 Soft Labeling 방식을 채택했다. 예를 들어, 10명의 평가자 중 8명이 'happy'로 판단하고, 2명이 'neutral'이라고 판단했다면, 해당 데이터의 레이블은 '[0.8, 0.2, 0, 0, 0, 0, 0]'으로 정의하는 방식을 사용했다.

2.2 멀티모달 모델

본 연구에서 제안하는 모델의 구조는 <그림 1>과 같다. 먼저 KoBERT[6]와 Mel-Spectrogram을 이용하여 Raw-텍스트와 Raw-오디오로부터 각각의 임베딩 벡터를 추출한다. 이후 오디오 임베딩은 2차원-Convolution을 거친 후 텍스트 임베딩과 병합된다. 이렇게 병합된 발화 임베딩을 시간순으로 8개씩 묶어 대화 임베딩을 생성한다. 대화 임베딩은 LSTM Layer에 입력되어 특징 벡터가 출력된다. 추출된 특징 벡터는 다시 Emotion Model과 Affective State Model에 각각 입력된다.

Emotion Model은 LSTM-Layer의 출력을 입력으로 사용하고, 7가지 감정에 대한 Softmax 확률을 출력한다. Affective State Model도 LSTM-Layer의 출력을 입력으로 사용하여 잠재 특징을 출력한다. 이후, Emotion Model에서 학습된 감정 특징을 Affective State Model의 잠재 특징과 병합한 후 각기 상이한 MLP Layer의 입력으로 사용하여 각성도 및 긍/부정도 값을 예측한다. 이때 모델의 예측값이 1과 5 사이의 연속적인 값이어야 하므로, Affective State Model의 출력에 $4 \times sigmoid(x) + 1$을 적용하여 최종 예측값으로 사용하였다.

2.4 손실 함수 (Loss Function)

Emotion Model의 손실 함수는 KL-Divergence Loss를 사용하였다. Affective State Model의 손실 함수로는 평균 제곱 오차(MSE)를 사용하였다.

대화의 감정(e), 각성도(a), 긍/부정도(v) 최적 예측값을 도출하기 위해 손실 함수 (1)를 제안한다. 수식(1)은 하이퍼 파라미터(Hyper Parameter) λ를 사용하여 감정 분류 손실과 각성도, 긍/부정도의 최소 제곱 오차 손실을 일정 비율로 합한다.

감정의 Class가 'neutral'이 대다수인 클래스 불균형 문제를 해결하기 위하여 손실 함수 (2)와 Class Balance[7] 수식(3)을 도입했다. 손실 함수 (2)는 불균형한 Class를 가진 감정을 최적하기 위해 Binary Cross Entropy Loss를 수식(3)으로 가중치를 주어 사용한다. 이때, n_y는 각 Class의 샘플 수이고 β는 하이퍼 파라미터로 0.99가 기본값으로 사용되었다.

$$(1)\ L(\hat{y}, y) = \lambda \cdot KLDiv(\hat{e}, e) + (1-\lambda) \cdot (MSE(\hat{a}, a) + MSE(\hat{v}, v))$$
$$(2)\ L(\hat{y}, y) = \lambda \cdot BCE(\hat{e}, e) + (1-\lambda) \cdot (MSE(\hat{a}, a) + MSE(\hat{v}, v))$$
$$(3)\ CB(\hat{y}, y) = \frac{1}{E_{n_y}} L(\hat{y}, y) = \frac{1-\beta}{1-\beta^{n_y}} L(\hat{y}, y)$$

3. 실험

우리는 모델의 학습과 평가를 두 개의 단계로 진행하였다. 우선 파라미터 평가를 위한 1차 실험과 실제 성능평가를 위한 2차 실험을 수행하였다. 1차 실험은 모델의 적합한 파라미터를 설정하기 위해 KEMdy19 데이터 세트에서 Session 13, 14, 15, 16을 검증 세트로 설정한 후 성능을 비교하였다. 이후 비교된 모델 중 가장 좋은 성능 보이는 파라미터를 선정하였다.

다음으로, 1차 실험에서 선정된 파라미터를 가지는 모델을 2차 실험에 사용하였다. 2차 실험은 공정한 평가를 위해, 선정된 파라미터 모델의 결과를 5-Folds 검증을 통해 최종 성능 지표로 제시하였다.

성능 지표에 표기된 Recall, Precision은 각각 감정분류 Recall, 감정분류 Precision을 나타낸다. Concordance Correlation Coefficient(CCC)는 예측정확도로, CCC(A)와 CCC(V)는 각각 각성도의 CCC, 긍/부정도의 CCC를 뜻한다.

3.1 제안 모델 성능

[표 1]은 멀티모달 데이터를 사용했을 때의 λ에 따른 Speaker 감정 예측 성능을 논멀티모달 데이터를 사용한 경우와 비교하여 보여준다. 멀티모달 데이터를 사용했을 때의 성능이 전반적으로 뛰어났으며 각 지표의 최고 성능은 굵은 글씨로 표시하였다.

[표 2]는 멀티모달 데이터를 사용했을 때의 λ에 따른 Listener 감정 예측 성능을 논멀티모달 데이터를 사용한 경우와 비교하여 나타낸다. Listener의 경우 멀티모달이 논멀티모달 보다 모든 경우에서 성능이 뛰어나지는 않았지만, 멀티모달 데이터를 사용했을 때 최고성능을 도출하였다.

Speaker와 Listener 예측 모델 모두에서 텍스트를 제외한 경우와 오디오를 제외한 경우를 비교했을때 눈에 띄는 성능저하를 보였다. 본 결과로부터 텍스트에 보다 풍부한 감정 정보가 포함되어 있음을 추론할 수 있다.

[표 1] Speaker에 대한 감정, 각성도, 긍/부정도 평가 결과

Input Data	λ	Recall	Precision	F1	CCC(A)	CCC(V)
Audio+Text	0.5	0.738	0.709	0.722	0.780	0.824
Audio+Text	0.66	0.748	0.719	0.733	0.745	0.860
Audio+Text	0.75	**0.759**	**0.731**	**0.744**	**0.791**	**0.869**
Audio+Text	0.8	0.696	0.670	0.682	0.783	0.803
Audio	0.66	0.489	0.466	0.477	0.540	0.588
Text	0.66	0.738	0.710	0.723	0.702	0.868

[표 2] Listener에 대한 감정, 각성도, 긍/부정도 평가 결과

Input Data	λ	Recall	Precision	F1	CCC(A)	CCC(V)
Audio+Text	0.5	0.712	0.683	0.696	0.746	**0.880**
Audio+Text	0.66	**0.744**	**0.713**	**0.728**	**0.756**	0.865
Audio+Text	0.75	0.740	0.710	0.724	0.712	0.870
Audio+Text	0.8	0.709	0.680	0.694	0.688	0.862
Audio	0.66	0.528	0.504	0.515	0.479	0.547
Text	0.66	0.728	0.697	0.711	0.716	0.861

3.2 감정이 각성도와 긍/부정도에 미치는 영향

[표 3]은 각성도, 긍/부정도에 감정을 병합하지 않고 학습 및 추론하였을 때의 성능을 보여준다. 이 경우 감정을 각성도와 긍/부정도에 병합하였을 때보다 낮은 F1 Score를 도출했다. 본 결과는 제안 모델이 감정 손실과 각성도 손실, 긍/부정도 손실에서 두 번의 규제를 가하므로 감정을 보다 정확히 예측할 수 있음을 나타낸다.

[표 3] 감정이 병합되지 않았을 때의 Speaker, Listener 성능

	Recall	Precision	F1	CCC(A)	CCC(V)
Speaker	0.719	0.690	0.704	0.751	0.827
Listener	0.722	0.692	0.706	0.689	0.876

3.4 Speaker, Listener 최고성능 5-Folds 검증

[표 4]은 Speaker에서 가장 높은 F1 Score를 보인 $\lambda = 0.75$을 사용한 제안 모델의 5-Folds 검증 결과를 나타낸다. [표 5]은 Listener에서 가장 높은 F1 Score를 보인 제안 모델에 CBLoss($\lambda = 0.9, \beta = 0.9$)를 사용한 모델의 성능 검증 결과이다.

[표 4] Speaker 모델의 5-Folds 검증 성능

	Recall	Precision	F1	CCC(A)	CCC(V)
Speaker_1	0.703	0.679	0.690	0.766	0.841
Speaker_2	0.731	0.707	0.718	0.766	0.824
Speaker_3	0.717	0.689	0.702	0.797	0.830
Speaker_4	0.759	0.731	0.744	0.791	0.869
Speaker_5	0.730	0.703	0.716	0.770	0.875
평균	0.728	0.702	0.714	0.778	0.848
표준편차	0.021	0.020	0.020	0.015	0.023

[표 5] Listener 모델의 5-Folds 검증 성능

	Recall	Precision	F1	CCC(A)	CCC(V)
Listener_1	0.642	0.617	0.629	0.754	0.842
Listener_2	0.720	0.698	0.709	0.717	0.814
Listener_3	0.650	0.625	0.636	0.626	0.778
Listener_4	0.745	0.715	0.729	0.723	0.877
Listener_5	0.713	0.685	0.698	0.735	0.852
평균	0.694	0.668	0.680	0.711	0.833
표준편차	0.045	0.044	0.045	0.050	0.038

4. 결론 및 향후 연구

본 연구는 대화의 맥락 정보와 멀티모달 데이터를 기반으로, 오디오와 텍스트 정보로부터 감정 정보를 예측하는 새로운 모델을 제안했다. 우리는 논멀티모달, 멀티모달 모델의 성능을 비교 평가했다. 또한 파라미터 탐색을 통해 성능 변화를 확인하고 가장 뛰어난 성능을 보이는 모델을 선택해 5-Folds 검증을 실시하였다. 제안된 멀티모달 모델은 텍스트와 오디오의 특징 벡터를 함께 사용함으로써 기존 연구보다 개선된 감정 분류 결과를 도출하였다. 하지만, 여전히 Data Imbalance 문제가 남아있어, Imbalance Sequential Data를 모델에 효과적으로 학습시키기 위한 방법론과 이를 감정분석 연구에 적용하는 시도가 필요하다.

참 고 문 헌

[1] C. N. Anagnostopoulos, T. Iliou and I. Giannoukos, “Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011”, Artificial Intelligence Review, Vol. 43(2), pp. 155-177, 2012.

[2] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar and T. Alhussain, "Speech Emotion Recognition Using Deep Learning Techniques: A Review”, *IEEE Access,* Vol. 7, pp. 117327-117345, 2019.

[3] A. J. AbdaouiAzé, S. Bringay and P. Poncelet, “Feel: a French expanded emotion lexicon.”, Lang Resour Eval, Vol. 51(3), pp. 833-855, 2017.

[4] J. Ma, H. Tang and W. L. Zheng, “Emotion Recognition using Multimodal Residual LSTM Network”, 27th ACM, pp. 176-183, 2019.

[5] K. J. Noh, C. Y. Jeong, J. Limm S. Chung and G. Kim, “Multi-Path and Group-Loss-Based Network for Speech Emotion Recognition in Multi-Domain Datasets”, Sensors, Vol. 21, 2021.

[6] Lee, Sangah, et al. "Kr-bert: A small-scale korean-specific language model." *arXiv preprint arXiv:2008.03979* (2020).

[7] C. Yin, M. Jia and T. Y. Lin, “Class-balanced loss based on effective number of samples.” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019.