멀티모달 감정인식을 위한 사전학습된 모델 기반 텍스트 및 음성 특징표현 활용 방법*

김준우○, 김동현, 도주성, 정호영
경북대학교 일반대학원 인공지능학과
kaen2891@gmail.com, kdh1477@gmail.com, jsean0423@gmail.com, hoyjung@knu.ac.kr

Strategies of utilizing pre-trained text and speech model-based feature representation for multi-modal emotion recognition

June-Woo Kim○, Dong-Hyun Kim, Ju-Seong Do, Ho-Young Jung
Department of Artificial Intelligence, Graduate School, Kyungpook National University

요 약

최근 딥러닝 기술의 발전에 따라 자연어 및 음성처리 도메인의 대규모 데이터셋으로 사전학습된 언어모델과 음향모델이 공개되어 널리 사용되고 있다. 멀티모달 감정인식 관점에서 사전학습된 모델의 효과성을 입증하기 위해 본 논문에서는 각각 한국어 및 영어 기반으로 학습된 언어모델 및 음향모델의 활용법을 탐구한다. 이를 위해 우리는 사전학습된 두 모달리티 기반의 모델로부터 멀티모달 감정인식을 위한 두 가지 특징표현 활용 방법에 대해 제안한다. KEMDy19, KEMDy20 데이터셋을 사용하여 진행된 멀티모달 감정인식 실험 결과는 제안된 방법이 단일 모달리티 기반의 결과보다 더 우수한 성능을 달성하였다.

1. 서 론

감정인식은 인간-로봇 상호 작용, 인간-컴퓨터 상호 작용 분야를 포함하지만 이에 국한되지 않는 광범위한 응용 가능성이 있는 연구 분야이다. AI 기반 시스템이 사용자의 입력 발화로부터 정확한 감정을 구별하는 것은 위 연구 분야의 주요 과제 중 하나이다. 특히, 대부분 상용화된 AI 시스템은 주로 사용자의 음성에만 집중하고 감정은 고려하지 않기 때문에 이에 해당하는 정확한 의도 분류가 불가능 할 수도 있다. 따라서, AI 시스템이 인간-컴퓨터 상호 작용이 되기 위해서는 인간의 정확한 감정과 그에 해당하는 의도를 분류하는 것이 중요하다.

감정을 인식하는 방법에는 이산적 및 연속적 표현(discrete and continuous representation) 두 가지 방법이 있다[1]. 이산적 접근 방식은 감정을 4-7개의 여러 범주로 분류하는 것으로 구성된다[2]. 대조적으로, 연속적 접근 방식은 각성(Arousal) 및 정서가(Valence) 차원으로 분리하여 측정하고, 이를 정서 분석에 활용한다[3]. 각성 차원은 각성(arousal)-수면(sleep), 활성(activation)-비활성(deactivation)과 같은 의미로 사용된다. 긍정, 부정 등 극성을 나타내는 정서가 차원은 쾌(pleasant)-불쾌(unpleasant) 차원과 동일한 의미로 사용된다[4-7]. 본 논문에서는 두 접근 방식 모두를 적용한다.

딥러닝 기반 멀티모달 감정인식은 서로 다른 두개 이상의 모달리티로부터 서로 상호 보완이 가능한 세련된 결과물을 내는 것을 목적으로 한다. [8]에서는 두 개의 동적 손실함수를 활용하여 감정과 나이 분류 성능을 개선하였고, [9]에서는 음성 데이터셋과 Google API를 사용하여 인식된 텍스트로 멀티모달 감정인식을 진행하여 성능을 개선하였다. [10]에서는 얼굴, 텍스트, 음성 모달리티를 이용하여 서로 증강 및 보완이 가능한 연구 결과를 보였다. 본 논문에서는 텍스트와 음성 정보 모두를 사용하여 멀티모달 감정인식 실험을 수행한다.

한편, 딥러닝 기술의 발전에 따라 자연어 및 음성처리 도메인에서 대규모 데이터셋 기반의 사전학습된(pre-trained) 모델들이 공개됨과 동시에 하위 분야(downstream)에서 적은 양의 데이터셋으로도 정확한 수준의 성능을 낼 수 있는 연구들이 많이 진행되고 있다. 특히, 비전사 데이터셋(unlabeled) 기반의 자기지도학습(self-supervised learning) 방법을 활용한 상위(upstream) 언어모델[11] 및 음향모델[12]을 활용하여 특정 도메인의 적은 데이터로 미세조정(fine-tuning) 학습을 한 모델의 성능은 기존 지도학습(supervised learning) 모델의 성능 대비 능가하였다.

이에 본 논문에서 우리는 사전학습된 언어모델[13]과 음향모델[12]을 활용하여 기존 텍스트 및 음성 데이터 각각의 유니모달 감정인식 성능 대비 더 나은 성능을 보이기 위해 두 모달리티를 활용한 두 가지의 요소별 특징표현 활용 방법을 제안한다. 사전학습된 언어모델은 한국어 기반의 KLUE-base 모델[13]을 사용하고, 사람의 음의 높이, 음색 및 감정을 판단하기 위한 음향모델은 영어로 학습된 wav2vec2.0 모델[12]을 활용한다.

* 이 논문은 2019년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (2019-0-00004, 준지도학습형 언어지능 원천기술 및 이에 기반한 외국인 지원용 한국어 튜터링 서비스 개발)

실험은 KEMDy19**, KEMDy20*** 데이터셋[14]으로부터 각각 이루어졌으며, 제안한 멀티모달 접근방식이 단일 모달리티 기반의 실험 결과 대비 각성 및 정서가를 평가하는 예측 정확도인 CCC(concordance correlation coefficient) 성능을 각각 텍스트에서 **1.11%** 및 **7.02%** 만큼 개선하였고, 음성에서는 **51.55%, 41.96%** 만큼 능가하였다.

2. 감정인식을 위한 특징표현 활용 방법

본 논문에서는 사전학습된 언어모델인 KLUE[13] 모델과 음향모델인 wav2vec2.0[12] 모델을 활용한다. 위의 두 모델 모두 대규모 비전사 데이터셋으로 사전학습 되었기 때문에 풍부한 표현(representation)을 보유하고 있으며, 이를 통해 적은 양의 제한된 데이터셋으로도 하위 작업에서 지도학습 기반 방법보다 더 나은 성능을 달성할 것으로 기대된다.

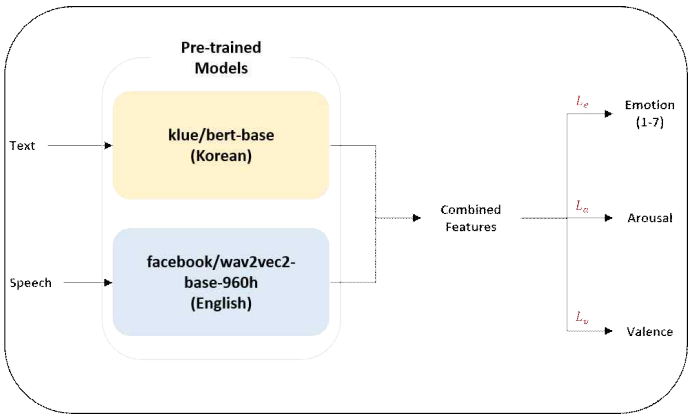

그림 1 사전학습된 단일 모달리티로부터 특징표현을 사용하는 제안된 멀티모달 감정인식을 위한 네트워크

그림 1은 본 논문에서 우리가 제안한 사전학습된 두 모달리티로부터 특징표현 활용을 통해 멀티모달 감정인식을 위한 네트워크이다. 두 사전학습된 모델은 모두 특징표현 추출기가 아닌 훈련가능한 상태(trainable)로 설정하고, 입력 데이터셋으로부터 마지막 은닉 상태(hidden state)를 출력으로 방출하도록 하였다. 이후 각 출력 개수에 비례한 선형 분류기를 추가하여 모델의 예측 값을 구할 수 있도록 하였다. 음향모델 및 언어모델은 각각 512, 768의 은닉 크기를 사용하였다.

2.1 요소별 결합 기반 특징표현 활용 방법

본 방법에서는 사전학습된 두 모델의 마지막 은닉 크기인 512(음성), 768(텍스트)로 마지막 타임 스텝의 출력 값을 결합한 뒤 선형 분류기를 통과하는 방법을 사용하였다. 해당 방법은 멀티모달 모델이 어떠한 데이터 손실 없이 기존 단일 모달리티가 보지 못하는 부분까지 특징표현을 추출할 것으로 기대된다. 이를 식으로 표현하면 다음과 같다.

$$X^i_{cat} = Concat\left(X^i_{text},\ X^i_{speech}\right) \in R^{(d_{text} + d_{speech})}$$

2.2 요소별 합 기반 특징표현 활용 방법

본 방법에서는 사전학습된 두 모델의 마지막 은닉 크기인 512(음성), 768(텍스트)의 크기를 통일시켜준 뒤 각 특징 값을 더하는 방법을 취하였다. 즉, 음성 특징표현 크기가 768으로 변경된 뒤 마지막 타임 스텝의 두 출력 값들을 요소별로 더하고(element wise-add), 선형 분류기를 추가하여 계산할 수 있도록 하였다. 해당 방법은 모델의 선형 분류기의 입력 크기가 변경되지 않고, 서로 다른 두 데이터를 동일한 은닉 크기로 스케일링 해주었기 때문에 첫 번째 방법 대비 더 안정적인 장점이 있다. 이를 식으로 표현하면 다음과 같다.

$$X^i_{Add} = Add\left(X^i_{text},\ FC\left(X^i_{speech}\right)\right) \in R^{d_{text}},$$
$$\text{Where } FC:\quad R^{d_{speech}} \mapsto R^{d_{text}}$$

3. 실험 및 결과

3.1 실험 세팅

본 논문의 모든 텍스트, 음성, 멀티모달 기반 감정인식 실험에 사용된 배치(batch) 크기는 32, 학습률(learning rate)은 1e-5, 에폭(epoch)은 5, 옵티마이저(optimizer)는 eps값이 1e-8인 AdamW를 사용하였다. 실험에 사용된 GPU는 A6000, 11.2의 CUDA 및 1.9.1의 Pytorch 버전을 사용하였다.

사전학습된 wav2vec 2.0 모델의 매개변수가 너무 큰 관계로 주어진 KEMDy19 및 KEMDy20 데이터셋의 모든 음성 데이터는 전부 5초로 샘플링하여 학습하였으며, 텍스트의 경우 사전학습된 KLUE 토크나이저[13]를 활용하여 모든 정보를 사용하였다.

KEMDy19 데이터셋의 경우 텍스트/음성의 매핑이 불가능한 12개의 파일을 제외하고 20,554개를 사용하였고, KEMDy20 데이터셋은 13,462개 모두를 사용하였다. 또한, 학습에 사용된 두 데이터셋 모두 5-fold 교차 검증이 진행되었으며, 실험 결과로부터 도출한 값들은 5개 결과의 평균값을 구하였다. 또한, 감정 레이블의 경우 복수의 정답이 존재하는 다중 레이블 분류 문제이기 때문에 아래의 식처럼 다중 레이블 일대일 손실값을 최적화하는 함수를 사용하였다.

$$-\frac{1}{C} * \sum_i y[i] * \log\left((1+\exp(-x[i]))^{-1}\right)$$
$$+ (1-y[i]) * \log\left(\frac{\exp(-x[i])}{1+\exp(-x[i])}\right)$$

본 논문에서 사용된 소스코드는 여기****에서 확인할 수 있다.

3.2 실험 결과

표1은 텍스트, 음성, 그리고 텍스트 및 음성 정보 모두를 함께 사용한 실험 결과를 보여준다.

KEMDy19 데이터셋의 F-score 관점에서 텍스트 기반 단일 모달리티의 성능 결과가 50.57%로 가장 좋았고, 음성 모달리티만 사용하였을 때 F-score와 긍/부정에 대한 가장 낮은 CCC 측정값을 획득하였다. 제안한 멀티모달 방법 중 요소별로 더한 방법이 텍스트 및 음성 모달리티 대비 각각 **1.22%, 51.66%** 능가하였다.

KEMDy20 데이터셋의 경우 모두 같은 F-score 값을 획

** KEMDy19: https://nanum.etri.re.kr/share/kjnoh/KEMDy19?lang=ko_KR

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

**** https://github.com/KNUAI/etri_multimodal

Data	Method	F-score (%)	CCC(%)		
			Arousal	Valence	Avg
KEMDy19	T	**50.57**	61.31	78.78	70.05
	S	43.86	31.25	7.96	19.61
	T+S (Cat)	49.93	**71.37**	70.71	71.04
	T+S (Add)	48.83	63.67	**78.86**	**71.27**
KEMDy20	T	85.18	44.86	61.07	52.97
	S		33.48	2.58	18.03
	T+S (Cat)		**58.15**	**67.01**	**62.58**
	T+S (Add)		54.21	60.59	57.4

표 1 KEMDy19 및 KEMDy20의 각 모달리티 및 멀티모달 별 F-score, CCC 결과(T: 텍스트, S: speech, T+S: 텍스트 및 음성 모두를 활용한 멀티모달, Cat: 요소별 결합, Add: 요소별 합)

득하였다. KEMDy19 데이터셋과 마찬가지로 음성 모달리티만 사용한 CCC의 값이 18.03%로 가장 낮았다. 해당 데이터셋에서도 제안한 멀티모달 방법 모두 단일 모달리티 성능들을 능가하였으며, 특히 요소별 결합한 방법으로부터 텍스트 및 음성 모달리티 대비 각각 **9.61%, 44.55%** 능가한 결과를 얻었다.

4. 결 론

본 논문에서 우리는 멀티모달 감정인식을 위해 널리 사용되는 사전학습된 언어모델 및 음향모델을 활용하여 각각의 모달리티 별 특징표현 활용 방법을 제안하였다. 제한된 감정인식 훈련셋인 KEMDy19 및 KEMDy20 데이터셋으로부터 본 논문에서 제안한 감정인식을 위한 멀티모달 훈련 방법을 활용하여 각성(Arousal) 및 정서가(Valence) 차원에서 기존 텍스트 모달리티 대비 4.07%, 음성 모달리티 대비 46.76% 만큼 개선하였다. 본 실험 결과를 통해 단일 모달리티 모델이 판단할 수 없는 부분을 멀티모달 모델이 서로 다른 모달리티를 상호 보완 및 개선된 판단을 해줌으로써 단일 모달리티 실험 결과 대비 능가하는 성능을 거둘 수 있었다.

참 고 문 헌

[1] Schuller, Björn W. "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends." *Communications of the ACM* 61.5 (2018): 90-99.

[2] Ekman, Paul, and Wallace V. Friesen. "Constants across cultures in the face and emotion." *Journal of personality and social psychology* 17.2 (1971): 124.

[3] Russell, James A. "A circumplex model of affect." *Journal of personality and social psychology* 39.6 (1980): 1161.

[4] Russell, James A. "Evidence of convergent validity on the dimensions of affect." *Journal of personality and social psychology* 36.10 (1978): 1152.

[5] Bradley, Margaret M., and Peter J. Lang. *Affective norms for English words (ANEW): Instruction manual and affective ratings*. Vol. 30. No. 1. Technical report C-1, the center for research in psychophysiology, University of Florida, 1999.

[6] Russell, James A., and Lisa Feldman Barrett. "Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant." *Journal of personality and social psychology* 76.5 (1999): 805.

[7] Hu, Minqing, and Bing Liu. "Mining and summarizing customer reviews." *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2004.

[8] Chae, Myungsu, et al. "End-to-end multimodal emotion and gender recognition with dynamic joint loss weights." *arXiv preprint arXiv:1809.00758* (2018).

[9] Xu, Haiyang, et al. "Learning alignment for multimodal emotion recognition from speech." *arXiv preprint arXiv:1909.05645* (2019).

[10] Mittal, Trisha, et al. "M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues." *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. No. 02. 2020.

[11] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

[12] Baevski, Alexei, et al. "wav2vec 2.0: A framework for self-supervised learning of speech representations." *Advances in Neural Information Processing Systems* 33 (2020): 12449-12460.

[13] Park, Sungjoon, et al. "Klue: Korean language understanding evaluation." *arXiv preprint arXiv:2105.09680* (2021).

[14] Noh, Kyoung Ju, et al. "Multi-path and group-loss-based network for speech emotion recognition in multi-domain datasets." *Sensors* 21.5 (2021): 1579.