대화 맥락과 사전학습 정보를 통한 멀티 모달 감정 인식 : 텍스트와 오디오를 중심으로

고혁훈[1], 주성호[1], 정교민[1,2]
[1]서울대학교 머신인텔리전스 연구실 [2]서울대학교 자동화 연구소
Hyukhunkoh-ai@snu.ac.kr seonghojoo@snu.ac.kr kjung@snu.ac.kr

Reflecting dialogue and pretrained information into Multi Modal Emotion Recognition: Focusing on Text and Audio

Hyukhun Koh[1], Seongho Joo[1], Kyomin Jung[1,2]
[1]Seoul National University Machine Intelligence Lab
[2]Automation and Systems Research Institute, Seoul National University
Hyukhunkoh-ai@snu.ac.kr seonghojoo@snu.ac.kr kjung@snu.ac.kr

요 약

본 연구에서는 딥러닝 기반 음성 및 자연어 모델을 이용하여 기쁨, 놀람, 분노, 중립, 공포, 슬픔의 7가지 감정을 판단해낼 수 있는 음성 텍스트 멀티 모달 모델을 만드는 것이 목표이다. 일상 속에서 사람들의 대화의 내용적인 면과 운율적인 면을 많이 반영한 KEMDy19[1]와 KEMDy20[2]을 활용하여 모델을 학습을 하였고, 그 과정에서 데이터 증강 기법을 이용하여 모델의 성능을 끌어 올렸다. 이러한 기법을 통해 한국어 사전학습 모델만이 아닌 영어권 사전학습 모델을 이용할 수 있게 되어서, SpeechT5와 wav2vec2.0 모델을 뼈대로 설정했다. 그 후 음성변환 및 음성인식 모델을 해당 Task에 적용하여 음성 및 텍스트 간의 연계성을 확보하였다. 그 결과 크게 향상된 정확도를 보였다.

1. 서 론

최근 인공지능 기술이 발전하면서, 한 가지 타입의 정보만 이용하는 모델보다 다양한 타입의 정보를 이용하는 모델이 대두되고 있다[1]. 이러한 기술적 경향은 특정 분야에 국한되지 않고 딥러닝의 큰 흐름이 되었다.

기존 텍스트 감정인식 분야에서는 딥러닝 기반 모델을 이용하여 사람의 감정을 파악할 때, 텍스트 정보만을 바탕으로 파악하는 것에는 한계가 있다[2], 요즘 딥러닝 기반 음성 모델 또한 언어모델을 결합해서 사용한다[3].

실제로 감정은 단일 요소로 결정되기보다는 복잡한 여러 요소들 간의 상호작용을 통해 표현된다. 따라서 다양한 모달리티를 고려한 감정인식 모델을 만드려는 시도가 있었다[4].

이러한 멀티 모달 모델들을 구축하는 경우, 사전학습된 초거대 모델이 필수적이다. 그러나 초거대 사전학습 모델의 경우, 대부분 많은 데이터 자원(High Resource language)이 구축되어 있는 미국에 많이 배포되어 있고. 적은 데이터 자원(Low Resource language)을 보유하고 있는 나라에서는 그 모델이 많지 않다. 또한 각각의 모달리티마다 독립적인 학습을 진행한 후 단순히 데이터를 더하거나 합쳐서 감정 라벨을 결정하는 것은 데이터 간의 상관성을 고려하지 못하는 경우가 많다.

따라서 본 연구에서는 음성-텍스트 멀티 모달 기반 프레임워크를 제시하여, 음성-텍스트 간 정보를 이용한 감정인식 모델을 구축하고자 한다. 또한 멀티 모달 모델을 위한 데이터 증강기법 및 학습 방법론을 고안하여 성능을 끌어올리고자 한다. 코드는 깃헙[3]에서 접근 가능하다.

2. 멀티 모달 감정분류 시스템

음성-텍스트 멀티 모달 모델에서 사전학습 모델을 사용하는 것에는 크게 두가지 이점이 있다. 첫 번째로, 데이터가 부족한 경우에 사전 학습 모델이 성능 향상에 유의미한 기여를 한다. 두 번째로, 일반화 성능이 올라간다. 즉, 보지 않은 데이터의 분포를 학습하지 않고도(Zero-shot setting) 해당 데이터의를 잘 예측할 수 있게 도와준다. 이러한 이점 때문에 사전학습된 모델을 사용하여 모델을 구축하는 것은 현대 딥러닝 기술에서 필수적인 방식이다.

음성-텍스트 멀티 모달 모델 중 공개된 대표적인 모델은 VALL-E[3], data2vec[5], wav2vec2.0[6],

[1] https://nanum.etri.re.kr/share/kjnoh/KEMDy19?lang=ko_KR
[2] https://nanum.etri.re.kr/share/kjnoh/KEMDy20?lang=ko_KR

[3] 깃헙코드링크

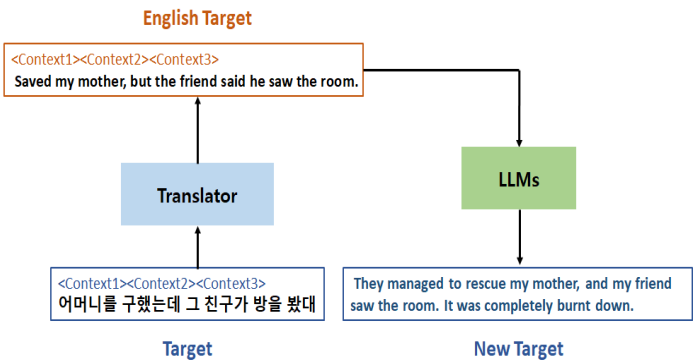
이 논문은 삼성전자, 2023년도 BK21 FOUR 정보기술 미래 인재 교육연구단 그리고 2023년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원 및 한국연구재단의 지원을 받아 수행된 연구임 [NO.2021-0-01343, 인공지능대학원지원(서울대학교) & NO.2021R1A2C2008855]

SpeechT5[7] 등이 있다. 그러나 한국어 사전학습 모델은 wav2vec2.0밖에 없기 때문에, 이러한 영어권 사전학습 모델들을 이용하기가 쉽지 않다. 본 연구에서는 이러한 한계점을 극복한 후, SpeechT5[5]에 meta-attention layer를 추가하여 모달리티간 얼라이먼트를 고려한 감정분류 시스템을 구축했다. 또한 텍스트와 음성 데이터셋을 엔트로피 및 거리기반 결합 학습방식을 이용하여 모델에 학습했으며, 그를 통해 멀티 모달 임베딩이 음성과 텍스트 간의 상호작용을 배울 수 있게 하였다. 그 후 완전일치방법(EM score)을 이용하여 모델을 평가하였다.

2.1 데이터 처리 및 증강

데이터셋의 label은 크게 valence와 arousal 및 감정라벨로 구성되어 있다. Valence와 arousal은 연속적인 데이터이고, 감정라벨은 이산적인 데이터이다. 본 연구에서는 Valence와 arousal 데이터는 리그레션 모델 기반으로 처리하였고, 감정라벨은 분류 모델 기반으로 다중 라벨적인 관점에서 접근하였다. 예를 들어, 하나의 음성-텍스트의 라벨이 [분노;기쁨]으로 동시에 되어있다면 각각 분노와 기쁨을 판별하도록 모델링을 하였다.

다음으로, 주어진 음성 스크립트 데이터들은 한국어로 되어 있기 때문에, 앞서 제시한 사전학습 모델을 이용하는 것에 어려움이 있다. 특히 언어 모델에서 거대 언어 모델들은(LLMs) 토큰아이저가 모델의 성능을 좌우하는데, 구축 방식에서 영어기반 BPE를 사용하였기 때문에 한국어를 처리하는 것 자체가 큰 장벽이 된다. 마지막으로, 해당 데이터셋이 대화 데이터셋이기 때문에, 해당 스크립트 이외에도 앞뒤의 음성-텍스트 데이터에 대화의 맥락이 담겨 있기 때문에 이를 고려해야 할 필요가 있다.



[그림 1] 대화맥락을 고려한 데이터 증강기법

이러한 요소들을 고려하여, 시중에 공개된 GPT-4[9]를 이용한 데이터 증강기법을 이용했다. GPT-4[9]의 번역 성능은 번역을 전문으로 하는 최신 모델들과 차이가 거의 없기 때문에[10], 이러한 텍스트 스크립트의 언어적 장벽 해소를 도와줄 수 있다. 또한 챗봇 분야에서 성능이 탁월하다고 알려져 있어, [그림 1]처럼 전후 맥락을 고려하면서 번역하기에 언어모델 관점에서 좋은 텍스트를 만들 수 있다. 이렇게 증강된 스크립트를 이용하여 우리는 기존에 사용할 수 없었던 영어로 사전학습된 모델을 이용할 수 있다.

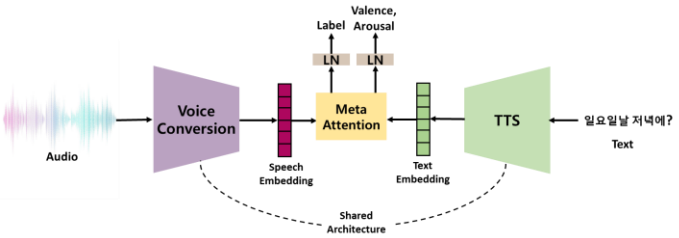
2.2 멀티 모달 모델

본 연구에서, 채택한 한국어 사전학습 모델로는 앞서 제시한 최신 모델들 중, wav2vec2.0만 한국어 모델이 존재해서 이를 채택했다[8].

또한 영어 사전학습 모델로는 SpeechT5[7]를 채택했다. 이 모델은 언어모델의 T5[11]에서 영감을 받아 구축된 통합된 음성 학습 모델이다. 우리 연구팀은 같은 감정 클래스에서 사람들이 공통적으로 보여주는 발화자의 억양과 같은 운율적 정보를 포함한 음성특징 정보를 추출하기 위해 음성 변환 모델의 인코더를 사용했다. 예를 들어 감정 중 '실망함'을 나타내는 오디오 그룹에서 사람들이 공유하는 음성적 특성이 있다고 가정한다. 또한 텍스트 정보를 가공하기 위해 음성 합성 모델의 인코더를 활용했다. SpeechT5는 음성변환 및 합성을 통합된 틀에서 학습한 모델이기에 추가적인 튜닝 없이 적용할 수 있고, 기존 wav2vec2.0과 Hubert와 같은 최신 모델들을 압도한다고 알려져 있기에, SpeechT5를 사용했다.

2.3 학습 방식

멀티 모달 감정 인식 분야에서 결합 학습방식(Jointly multi-task)이 탁월하다고 알려져 있다[12]. 따라서 본 연구에서, 사전학습모델을 고정하고 meta-attention layer와 각각 라벨에 맞는 선형레이어(linear layer)를 추가하여 동시 학습을 진행하였다.



[그림 2] 음성-텍스트 멀티 모달 모델 구조

[그림 2]에서 주어진 사전학습 모델을 샴(Simaese) 네트워크로 초기화 하여, 각각 독립적으로 미세조정을 하였다. 이 경우 각각의 독립된 모델은 동일한 구조를 지니고 있다. 그 뒤 각 모델의 입력으로 텍스트와 음성을 동시에 넣고, 결합 학습을 하였다. 학습을 할 경우에, 모델의 출력에 두 개의 선형레이어를 두어서, 하나의 레이어는 감정라벨 기반으로 크로스 엔트로피기반의 분류학습을 하였고 각각의 감정마다 분류기 두었다. 다른 하나의 선형레이어는 valence와 arousal 라벨을 기반으로 MSE 학습을 보조적으로 진행하였다.

2.4 성능평가 방식

성능을 평가하는 척도는 요소별 Accuracy 대신에 Exact Match Score 방식을 채택했다. Multi-label 문제로 해당 task를 해석하였기에, 단순 개별요소로 Accuracy를 재는 것은 실제 성능과는 품질차이가 존재한다. 따라서 multi-

label이 정확히 일치하는 것을 기준으로 측정하였다.

3. 성능평가

본 연구에서는 크게 두 가지 관점에서 성능 실험을 하였다. 첫 번째로, LLMs 기반 데이터 증강 기법을 사용하여 타 언어권 국가의 사전학습 모델을 사용하는 것이 우리나라의 사전학습 모델에 비해 유의미한 지 탐색해보는 것이다. 두 번째로, 실제로 감정분석을 하는 과정에 있어, 딥러닝 모델적으로 단일 모달리티에 비해 멀티 모달리티를 사용했을 경우 어떠한 이득이 있는 가를 측정해 보았다. 실험 세팅의 모든 경우에서 모델이 30 에포크 이내에서 수렴하는 것을 관측하였기에, 에포크를 30으로 설정하였다. 번외로, 본 연구에 주요 시사점이 아니기에 실험한 것을 올리지 않았지만, valence 및 arousal 정보와 감정 라벨을 동시에 학습하는 것 또한 도움이 됨을 알 수 있었다.

		EM	Distance
Model	SpeechT5 (VC only)	82.69	0.07
	SpeechT5 (TTS only)	79.81	0.12
	k-wav2vec2.0	84.53	0.178
	SpeechT5 (multi-modal)	**89.58**	**0.068**

[표 1] 모델 성능 비교표

[표 1]에서 EM은 Exact Match를 의미하고, Distance는 Mean Squared Error이다. VC는 Voice Conversion을 의미하고, TTS는 Text-to-Speech이다. 결과적으로 Text 정보만을 고려한 것은 눈에 띄게 성능저하를 가져오는 것을 확인할 수 있다. 음성 정보를 고려하는 것은 Text정보에 비해 성능이 좋긴 하나, 멀티 모달리티를 고려한 것에 비해 성능이 많이 뒤쳐지는 것을 확인할 수 있었다. 또한 본 연구의 방법론을 따라 멀티 모달 모델을 학습을 했을 경우에 높은 성능을 얻을 수 있었기에, 감정인식 시스템에서 멀티 모달을 고려하는 것의 중요성을 알 수 있었다. 또한 국내 사전학습 모델을 사용하지 않고도, 최고 성능을 내는 것을 알 수 있었다.

4. 결론 및 향후 과제

본 연구에서는 데이터셋에서 음성과 텍스트 정보만을 이용한 모델을 구축하였다. 그 결과 하나의 모달리티만 이용하는 경우보다 월등한 성능의 증가가 있었고, 이러한 증강기법을 통해, 사전학습 모델을 이용하는 관점에서 언어 장벽을 해소할 수 있음을 보였다. 앞으로 다른 연구자들이 이용할 수 있는 사전학습 모델의 범위가 넓어질 것이다.

본 연구에서는 각각의 모듈을 미세조정 후에 상위 레이어에서 상호작용을 고려하는 연구를 진행하였다. 그러나 이러한 방법은 데이터 간 상호작용을 완벽하게 고려하지 못하기에, 추후 자기지도 학습 방법론을 도입하여 데이터 간의 분포를 직접적으로 배우도록 유도해보고자 한다. 또한 input으로 각각 모듈에 따로 들어가는 것에서 함께 들어갈 수 있도록 토큰아이저를 새롭게 디자인 후에 모델 파라미터 수와 비용 측면에서 효율성을 확보해 볼 것이다. 마지막으로 EDC와 같은 기타 데이터 기존 모델에 추가적으로 고려하여 개선된 모델을 만들어볼 것이다.

참고문헌

[1] Wang, Xiao, et al. "Large-scale multi-modal pre-trained models: A comprehensive survey." arXiv preprint arXiv:2302.10035 (2023).

[2] Alvarez-Gonzalez, Nurudin, Andreas Kaltenbrunner, and Vicenç Gómez. "Uncovering the limits of text-based emotion detection." arXiv preprint arXiv:2109.01900 (2021).

[3] Wang, Chengyi, et al. "Neural Codec Language Models are Zero-Shot Text to Speech Synthesizers." arXiv preprint arXiv:2301.02111 (2023).

[4] 이서연, 김원중, “멀티 모달 데이터를 활용한 결정 융합 기반 감정 분류 모델” 2022년 한국컴퓨터종합학술대회 논문집 VOL 49 NO. 01 PP. 2285 ~ 2287 (2022. 06)

[5] Baevski, Alexei, et al. "Data2vec: A general framework for self-supervised learning in speech, vision and language." International Conference on Machine Learning. PMLR, 2022.

[6] 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.

[7] SpeechT5: Unified-Modal Encoder-Decoder Pre-Training for Spoken Language Processing (Ao et al., ACL 2022)

[8] K-Wav2vec 2.0: Automatic Speech Recognition based on Joint Decoding of Graphemes and Syllables (Jounghee et al, arxiv)

[9] OpenAI, “GPT-4 Technical Report”, arxiv preprint arxiv: 2303.08774(2023)

[10] Zhu, Wenhao, et al. "Multilingual Machine Translation with Large Language Models: Empirical Results and Analysis." arXiv preprint arXiv:2304.04675 (2023).

[11] Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." The Journal of Machine Learning Research 21.1 (2020): 5485-5551.

[12] R. Peri, S. Parthasarathy, C. Bradshaw and S. Sundaram, "Disentanglement for Audio-Visual Emotion Recognition Using Multitask Setup," ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 2021, pp. 6344-6348, doi: 10.1109/ICASSP39728.2021.9414705.