MLP-Mixer 구조를 활용한 대화에서의 멀티모달 감정 인식*

방나모‡ 연희연○‡ 이지현‡ 구명완‡‡
서강대학교
email: {namo950815,yeen214, jhlee22, mwkoo}@sogang.ac.kr

MMM: Multi-modal Emotion Recognition in conversation with MLP-Mixer

Na-Mo Bang‡ Heui-Yeen Yeen○‡, Jee-Hyun Lee‡, Myoung-Wan Koo‡‡

요 약

최근 인공지능(AI) 분야에서 멀티모달 데이터를 활용해 비언어적•언어적 정보를 함께 반영하는 연구가 활발히 진행되고 있다. 감정 인식 분야에서도 사용자의 발화와 음성 톤 등에 기반한 발화 이해를 통해 감정 인식 성능을 높일 수 있다는 점에서 멀티모달의 중요성이 커지고 있다. 따라서 본 연구에서는 한국어 기반 멀티모달 감정 데이터셋(KEMDy20)의 발화 음성(Speech) 및 문맥적 의미(Text) 데이터를 결합하여 학습하는 MLP-Mixer방식의 Multi-modal MLP Mixer 방법론을 제안한다. 그 결과, 감정인식 Task에서 두 가지 모달리티 데이터를 단순 결합했을 때보다 Weighted F1 score 기준 8.1 정도 향상되었으며, 최근 제안되고 있는 멀티모달 결합 방식에 비해 훨씬 적은 파라미터로도 높은 성능을 냄을 보였다. 이를 통해 방대한 양의 멀티모달 데이터의 학습을 효율적으로 처리할 수 있도록 하여, 실제 서비스에 인공지능 모델 적용이 용이할 수 있도록 한다.

1. 서 론

인공지능의 급속한 성장과 함께 멀티모달 감정 인식은 주로 대화 생성, 사용자 의도 이해, 멀티모달 상호 작용 등과 같은 많은 어려운 작업에서 잠재적인 응용을 위해 주요 연구 주제가 되고 있다.

이에 따라 대화형 감정 인식 시스템을 사용하여 사용자 감정을 분석하여 적절한 응답을 생성하기 위해 오디오, 시각 및 텍스트 양식을 사용하는 멀티모달 데이터 세트가 등장했다. 멀티모달 정보를 활용하기 위해 모달리티 임베딩 값을 단순히 결합하는 방식부터 최근에는 모달리티의 특징 값들을 Cross-attention 방식으로 결합하는 방식이 State-of-the-art의 성능을 달성하고 있다. 하지만 이러한 Attention 방식의 결합은 복잡한 연산 방식과 원시 데이터로부터의 학습에 의존하는 경향이 크다. 따라서 본 연구에서는 각 모달리티 별 인코더에서 나온 임베딩 값을 MLP-Mixer 구조를 활용하여 결합한다. Attention을 활용하지 않고 퍼셉트론 layer에 Mixer 구조를 적용함으로 기본적인 행렬 곱셈과 데이터 레이아웃 변경(reshape와 전치), 스칼라 비선형성만 고려할 수 있게 한다.

따라서 단순한 연산 방식에도 불구하고 MLP-Mixer 기반 Multi-modal MLP Mixer(MMM)가 경쟁력 있는 결과를 가짐을 한국어 대규모 멀티모달 대화 데이터 셋(KEMDy20; Korean Emotional Multi-modal Dataset in 2020)**을 통해 보인다. 이로써 단순 모달리티 데이터 셋보다 방대한 멀티모달 데이터 셋에서 각 모달리티의 정보 값을 효율적으로 결합하여 비언어적•언어적 정보를 통합한 최적의 결과를 도출할 수 있도록 하였다.

2. 관련 연구

현재까지의 연구를 살펴보면 대화에서의 감정 인식 성능을 위해 멀티모달 데이터로 학습을 시킬 때 시각 데이터와 텍스트 데이터의 결합이 주를 이루었다. 또한 결합 방식에서 단순 결합 방식이나 Attention 방식을 적용하는 연구가 제안된 바 있다.

2.1 MLP-Mixer

컴퓨터 비전 분야에서 처음으로 제안된 방식이다. 비전 분야에서 강력한 성능을 보이던 CNN을 뛰어넘은 self-attention layer 기반의 트랜스포머 모델이 최첨단 성능을 달성하였다. 하지만 보다 단순한 연산으로 동일한 성능을 내는 여러 층의 퍼셉트론을 반복적으로 적용한 MLP-Mixer 구조가 제안되었다[1]. 이후 각 모달리티 정보에 대해 학습의 효율성을 올리는 연구는 많이 진행되었으나, 멀티모달 데이터를 MLP-Mixer 방식으로 감정인식 Task에 적용한 연구는 진행된 바 없다.

2.2 음성과 텍스트를 이용한 감정 인식

음성 데이터를 더 잘 이해하기 위해 텍스트 데이터와 오디오 신호를 동시에 활용하는 새로운 심층 이중 반복(RNN) 인코더를 제안한 연구가 있다[2]. 감정 대화는 음성과 텍스트로 구성되므로, 이중 반복 신경망을 사용하여 오디오와 텍스트 시퀀스의 정보를 인코딩한 다음 해당 정보들을 결합하여 감정 클래스를 예측하는 방법론을 제시하였다. 하지만 복잡한 연산 과정과 많은 파라미터를 가지며 학습이 진행되는 비효율적인 측면이 있다.

*이 논문은 2022년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2022-0-00621,대화 기반 설명가능성을 멀티모달로 제공하는 인공지능 기술 개발)

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

‡공동 주저자로 논문에 대한 기여도가 같음. ‡‡교신저자

3. 연구 방법

3.1 데이터셋

3.1.1 KEMDy20

한국어 기반 멀티모달 감정 데이터셋(KEMDy20; Korean Emotional Multi-modal Dataset in 2020)은 발화 음성, 발화의 문맥적 의미(Text) 및 생리반응 신호- 피부전도도(EDA-electrodermal activity), 맥박관련 데이터(IBI-Inter-Beat-Interval), 손목 피부온도와 발화자의 감정과의 연관성 분석을 위해 수집한 멀티모달 감정 데이터셋이다. 감정 레이블은 7개 (기쁨, 놀람, 분노, 중립, 혐오, 공포, 슬픔)로 구성되어있다. 대화 전체 레벨이 아닌 각 발화 별로 감정 어노테이션이 되어있다[3].

3.1.2 한국어 감성 대화 말뭉치

KEMDy20 데이터셋의 “중립” 감정 레이블에 치우친 데이터 불균형 문제를 보안하기 위해 AI Hub에서 감성 대화 데이터셋의 음성 데이터 10,000문장과 이를 전사한 텍스트 데이터를 활용하였다. 감정 레이블은 6개의 대분류(분노, 슬픔, 불안, 상처, 당황, 기쁨)로 되어있으며, KEMDy20 데이터셋 레이블에 맞게 (기쁨-기쁨, 놀람-당황, 분노-분노, 혐오-상처, 공포-불안, 슬픔-슬픔)으로 맵핑 하였다.

3.1.3 최종 데이터

최종적으로 학습 및 평가에 사용된 데이터는 표 1과 같다. 데이터셋은 세션(동일 그룹), 스트립트(동일 주제), 발화 레벨로 나뉘며 세션을 기준으로, 80%를 학습 데이터셋으로, 나머지 20% 데이터셋을 평가에 활용했다.

구분	Train	Test	총계
세션 레벨	32	8	40
스크립트 레벨	8192	2048	10240
발화 레벨	19086	5020	24106

표 1 최종 데이터셋 구성

그림 1를 보면, KEMDy20 데이터셋의 감정 라벨과 한국어 감성 대화 말뭉치를 합쳤을 경우 중립 라벨을 제외한 나머지 감정 라벨 분포가 균등함을 확인할 수 있다.

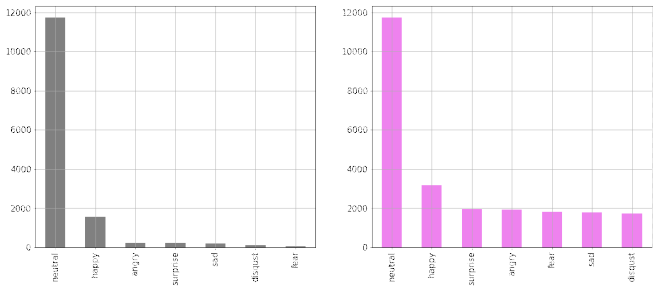

그림 1 증강을 통한 데이터 불균형 완화 전후 분포 비교

3.2 Multi-modal MLP Mixer

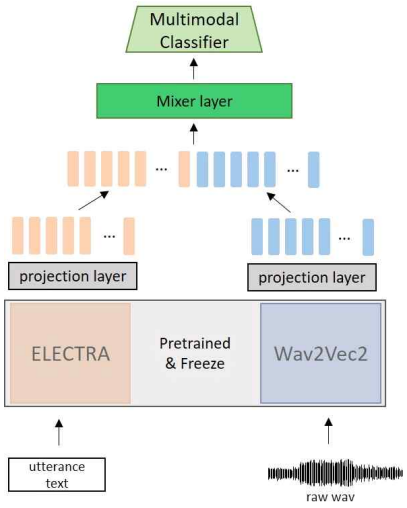

그림 2 Multi-modal MLP Mixer(MMM) 모델 구조

본 연구에서 제안하는 멀티 모달 아키텍처 구조는 그림 2와 같다. 멀티 모달 데이터 (음성, 텍스트)의 특징 정보들을 융합하기 위해, Pre-trained된 Wav2vec2.0[4] 기반 음성 모델과 KcELECTRA[5] 기반 텍스트 모델을 이용했다. Wav2vec 2.0의 히든 스테이트와 KcELECTRA의 히든 스테이트를 연결하는 방식을 취했고, 각 모달리티 데이터를 학습하는 중에는 다른 모델을 freeze해서 서로 방해받지 않도록 했다. 그리고 연결 이후 마지막 최종 분류 층에서 드롭아웃과 GELU를 적용해 성능을 높였다.

본 연구에서 제안하는 멀티 모달 특징 값 결합방식 MLP-Mixer는 그림 3과 같다. 각 모달리티 인코더에서 나온 히든 스테이트 값에 projection layer를 통과시켜 나온 각 벡터 값을 융합한 뒤 1개의 MLP-Mixer layer를 거쳤다. MLP-Mixer는 시퀀스 결합 블록(각 모달리티의 시퀀스 방향으로 layer normalization을 진행한 후 전치시킨 다음 MLP를 지남)과 모달리티 결합 블록(원래대로 전치시킨 다음 MLP를 거침)으로 이루어져 있다. 따라서 Position embedding 없이도 시간적 정보를 기억하기 때문에 복잡한 연산 과정 없이 MLP만으로도 멀티 모달 정보를 학습할 수 있다.

4. 연구 결과

성능 평가는 레이블이 불균형한 데이터에서의 성능 평가 지표로 사용되는 “weighted F1-Score”을 평가지표로 삼았다. 동일한 Task에서 기존의 방법론과의 성능 비교 및 모델 파라미터 수 비교를 통해 학습의 효율성을 검증하였다.

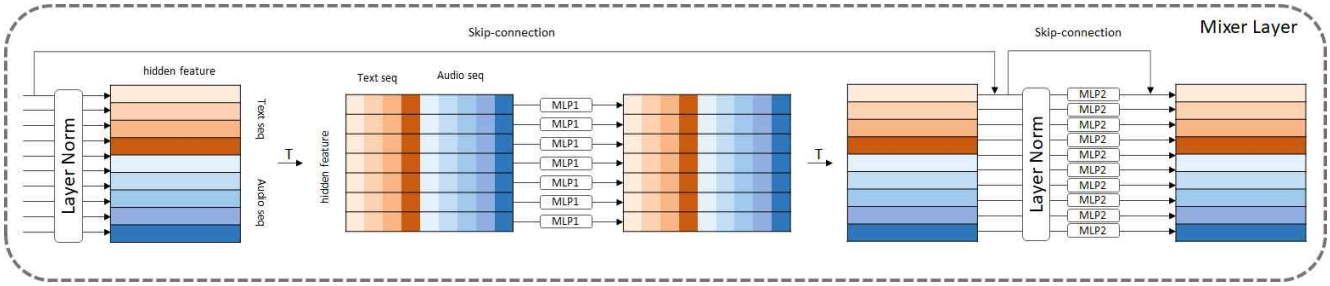

그림 3 MLP-Mixer 구조도

4.1 방법론 별 성능 분석

기존 연구에서 제안된 모달리티 특징 값을 합치는 4가지 방법론과 함께 비교를 진행하였다. (1) Concatenation : 각 모달리티 인코더에서 나온 임베딩 값을 단순히 이어 붙이는 방식 (2) 3-way : 각 모달리티 인코더에 나온 임베딩 값을 단순 융합, 행렬 요소 차이에 절댓값을 취한 연산, 곱셈 연산 (element-wise product) 방식으로 결합하는 방식[6] (3) Cross-attention: 각 모달리티 인코더에서 나온 임베딩 값을 트랜스포머 블럭으로 Cross-attention을 취하는 방식이다.

방법론	Accuracy	Weighted Precision	Weighted F1 Score
(1) Concat	71.75	69.58	70.25
(2) 3-way	73.12	72.08	72.13
(3) Cross-attention	77.74	78.29	77.34
Ours	**78.88**	**78.45**	**78.41**

표 2 특징 혼합 방식별 30epoch 학습 시 성능 비교

동일한 epoch를 기준으로 성능은 표 2과 같으며 MLP-Mixer 구조가 단순함에도 멀티모달 각 특징 값을 잘 결합하여 최적의 성능을 낸다는 것을 보인다.

4.2 학습의 효율성 분석

최근 멀티모달 특징 값 결합 방식의 state-of-the-art를 달성하고 있는 Cross-attention 방식과 동일한 레이어 개수를 사용하여 학습을 진행했을 때 표 3을 보면 약 10배 정도 적은 모델 파라미터를 가짐을 확인할 수 있다.

방법론	사용 레이어	학습 파라미터
Transformer (Cross Attention)	1	15758855
MLP Mixier	1	**1547527**

표 3 Cross-attention 방식과 모델 파라미터 개수 비교

표 2를 보면 동일한 epoch에서 MLP-Mixer 방식이 훨씬 적은 파라미터로도 Cross-attention 방식보다 성능 대비 좋은 효율성을 보였다.

5. 결론

본 연구는 대화에서의 감정인식을 위해 비언어적•언어적 정보를 함께 반영할 수 있도록 음성 (Speech) 및 문맥적 의미(Text) 데이터를 결합하여 학습하는 MLP-Mixer방식의 Multi-modal MLP Mixer 방법론을 제안하였다. 기존에 제안된 멀티모달 특징 값 결합 방식보다 현저히 적은 모델 파라미터 개수에도 불구하고 높은 성능을 낼 수 있다는 것을 보였다. 이는 인공지능 모델의 학습을 효율적으로 만들어 실제 사용에도 더욱 용이할 수 있도록 한다.

향후에는 대화 문맥을 함께 결합하여 대화에서의 감정 변화를 추적할 수 있도록 하고, 생체 정보를 비롯한 다양한 멀티모달 특징 값을 결합하여 복합적 멀티모달 대화 감정 인식 시스템을 구현하고자 한다.

참고문헌

[1] Tolstikhin, Ilya O., et al. "Mlp-mixer: An all-mlp architecture for vision." Advances in Neural Information Processing Systems 34 (2021).

[2] Lee, Yoonhyung, Seunghyun Yoon, and Kyomin Jung. "Multimodal Speech Emotion Recognition Using Cross Attention with Aligned Audio and Text." INTERSPEECH. (2020).

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

[4] Baevski, Alexei, et al. “wav2vec 2.0: A framework for self-supervised learning of speech representations.” Advances in Neural Information Processing Systems 33 (2020).

[5] Junbum Lee, “KcELECTRA: Korean comments ELECTRA” (2021).

[6] Conneau, Alexis, et al. "Supervised learning of universal sentence representations from natural language inference data." arXiv preprint arXiv:1705.02364 (2017).