Residual Self-Attention Cross Fusion Network (RSA-CFN): 상호작용에 기반한 멀티모달 아키텍처 퓨전 연구

이현호, 임영진○, 배지원
카카오뱅크, SK C&C, 무소속
gusgh1510@gmail.com, zerojin0502@gmail.com, pjt.jw101@gmail.com

Residual Self-Attention Cross Fusion Network (RSA-CFN): Multimodal Architecture Fusion based on Interaction

Hyeon-Ho Lee, Young-Jin Lim, Ji-Won Bae
KakaoBank, SK C&C, Independent Scholar

요 약

본 연구에서는 한국어 기반 멀티모달 감정 데이터셋(KEMDy20)를 활용하여, 발화자의 음성(Speech) 및 대화 정보(Text)를 결합해 감정을 분류하는 Architecture Fusion 방법론을 적용해 보고 성능을 개선하는 모델을 설계했다. Architecture Fusion은 여러 딥러닝 모델을 결합하여 하나의 모델을 생성하고 각 모델의 특징을 활용하여 보다 효과적인 학습을 수행하는 방법이다. 본 연구에서는 Early Fusion, Late Fusion의 모델의 성능을 비교했다. 이후, 기존 방식보다 유의미한 성능 향상을 내는 4가지 모델 구조를 설계했고, 모달리티 간의 상호 관계를 고려하는 Residual Self-Attention Cross Fusion Network (RSA-CFN)를 제안한다. 해당 모델의 성능은 Weighted F1-Score가 89.3%로 선행 연구(MLP-Mixer)[1] 대비 우수한 성능을 보여줬다.

1. 서 론

감정 상태를 분석하는 데는 다양한 정보가 필요하다. 본 연구에서는 KEMDy20 데이터[2]를 활용하여, 대화 상황에서 감정을 7가지로 분류하는 멀티모달 모델링을 진행했다. 연구의 주요 내용은 다음과 같다.

1) Early Fusion과 Late Fusion 성능 비교를 통해, 효과적인 Architecture Fusion 방법론을 실험한다.
2) 기존 Architecture Fusion을 개선한 다양한 모델을 설계하고 성능을 비교한다.
3) Self-Attention을 통해 모달리티 간 상호작용을 살펴본다.

2. 관련 연구

멀티모달 방법론에 대한 연구는 최근 ERC 분야에서 핵심적인 관심사로 대두되고 있다. 다양한 모달리티 정보들을 결합하여 하나의 분석 결과를 도출하면 보다 정확한 감정 분석이 가능하기 때문이다. 이에 다음과 같은 다양한 선행 연구가 제안된 바 있다.

이미지와 텍스트 멀티모달 분야에서는 모달리티 내 관계와 모달리티 간 관계를 통합된 심층 모델에서 공동으로 모델링하여 이미지와 문장 매칭을 위한 MultiModality Cross Attention Network를 제안했다.[3] 해당 모델에서는 각 모달리티 내 관계뿐만 아니라 이미지 영역과 문장 단어 간의 모달리티 간 관계를 활용하여 이미지와 문장 매칭을 위해 서로를 보완하고 향상할 수 있다.

다른 연구에서는 여러 계층에서 'Fusion Bottlenecks'를 사용하는 트랜스포머 기반의 아키텍처를 개발했다.[4] 이는 기존의 Self-Attention과 비교하여 서로 다른 모달리티 간의 정보가 소수의 병목 지점을 통과하도록 하여 각 모달리티에서 관련 정보를 수집 및 압축하고 필요한 정보를 공유하도록 유도했다.

3. 연구 방법

본 논문에서는 일반인 대상 자유 발화 데이터셋인 KEMDy20을 사용했다. 해당 데이터는 7개의 단일감정(중립, 기쁨, 화남, 놀람, 슬픔, 불안, 혐오)의 조합인 24개의 복합감정 라벨로 이루어져 있으며, 중립이 총 82%를 차지하는 불균형 데이터다. 이러한 불균형 문제를 완화하고자 층화 추출법과 데이터 증강을 사용했다.

3.1 전처리 및 특징 추출

3.1.1 전처리 (층화 추출법)

성능평가 시 전체 데이터셋의 80%에 해당하는 데이터를 학습에 사용하고 20%의 데이터를 평가에 사용했다. 데이터 전처리 이전 학습데이터와 평가데이터를 분할함으로써 데이터 처리에 따른 평가 지표 왜곡이 이루어지지 않도록 했다. 또한 데이터를 표본 추출하는 방법으로는 층화 추출법을 사용하여, 모델에 대한 학습 및 평가 신뢰도를 높였다.

3.1.2 전처리 (데이터 증강)

2개 이상의 복합감정 데이터의 경우, 중립과 함께 존재하는 2가지의 복합감정을 중립 외 감정으로 치환하는 Label Transform을 수행했다. 특히 학습 데이터에 대해서만 전처리하여 데이터 손실을 최소화하면서도 평가 데이터의 실제 분포를 잃지 않고자 했다. 전처리 이후 학습 데이터는 10,874건으로, 전처리 이전 10,262건 대비 612건의 데이터 증강이 이루어졌다.

표 1 Label Transform에 따른 학습 데이터 분포 변화

	학습 데이터		평가 데이터
	전처리 이전	전처리 이후	
Neutral	8,896(86.69%)	8,896(81.81%)	2,224(86.67%)

Happy	946(9.22%)	1291(11.87%)	237(9.24%)
Angry	115(1.12%)	198(1.82%)	29(1.13%)
Surprise	125(1.22%)	181(1.66%)	31(1.21%)
Sad	97(0.95%)	168(1.54%)	24(0.94%)
Fear	34(0.33%)	51(0.47%)	9(0.35%)
Disgust	49(0.48%)	89(0.82%)	12(0.47%)
총계	10,262	10,874	2,566

3.1.3 특징 추출

Fusion 방법론 적용 이전, 음성 및 텍스트 모달리티 간 특징 추출을 위해 사전학습모델을 사용했다. 음성에 대한 사전학습 모델은 현재 SOTA 모델인 Facebook의 wav2vec 2.0[5]를 사용했다. 텍스트 데이터에는 RoBERTa 계열 중 다국어 모델인 XLM-RoBERTa와 한국어 특화 모델인 KLUE-RoBERTa를 사용했다.[6] 제한된 학습 환경과 실험에서 효율성을 위해 Base를 기본으로 진행했으나, 음성 데이터에서 Base 모델은 과적합 되어 Large 모델을 사용했다.

3.2 Early Fusion & Late Fusion

본 연구에서 활용하고 있는 한국어 대화 데이터셋에 대한 Fusion 방법론별 성능을 살펴봤다. 학습을 위한 손실함수로는 Cross-Entropy Loss를 사용했다. 성능 평가의 경우 불균형 데이터에서 평가 지표로 쓰이는 Weighted F1-Score를 사용했다.

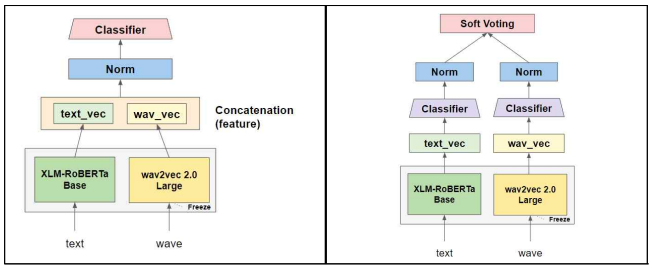

그림 1 Early Fusion 구조(좌)와 Late Fusion 구조(우)

Early Fusion은 서로 다른 모달리티의 데이터를 결합하여 하나의 특징 벡터(Feature Vector)로 만든 후, 이를 모델의 입력으로 사용해 여러 모달리티를 동시에 고려하여 학습하는 장점이 있고, Late Fusion은 모달리티별 모델의 예측 결과를 결합하는 방식으로 모달리티별 고유 특성에 대한 정보를 보존하며, 모델의 설계가 유연하다는 장점이 있다.

표 2 Architecture Fusion 방법론별 모델 성능

	Early Fusion	Late Fusion
F1-Score(W)	**0.89021**	0.8882

한국어 대화 데이터셋에서는 Early Fusion이 Late Fusion보다 더 좋은 성능을 가지는 것으로 나타났다. Early Fusion은 다양한 모달리티의 정보를 학습 이전에 결합하므로 각 모달리티의 상호작용을 좀 더 효과적으로 학습할 수 있지만, Late Fusion은 각 모달리티에서의 특징을 독립적으로 학습한 후 이를 결합하므로 상호작용에 대한 학습이 상대적으로 부족하다. 이에 본 연구에서는 Early Fusion을 기반으로 좀 더 효과적인 모델을 탐색하고 개발했다.

3.3 Model Architecture 설계

Stack 모델은 단일 분류기를 통해 예측하는 것이 아니라 다중 분류기들의 예측값을 활용하는 구조로 설계했고, Residual 모델은 Skip Connection을 통해 입력값을 출력값에 더해 활용하는 구조로 설계했다. Residual Self-Attention (이하 RSA) 모델은 각 모달리티간의 상호작용과 중요한 요인을 추출하기 위해 Self-Attention을 사용했고 이후 구조는 Residual과 동일한 방식으로 설계했다.

Residual Self-Attention Cross Fusion Network(이하 RAS-CFN) 모델은 각 모달리티의 상호작용을 학습시키기 위해 외적곱 형태의 Fusion 방식으로 설계했고, 이는 모달리티 간 상호작용을 강화하고 상호보완적 특성을 추출하는 효과를 보였다.

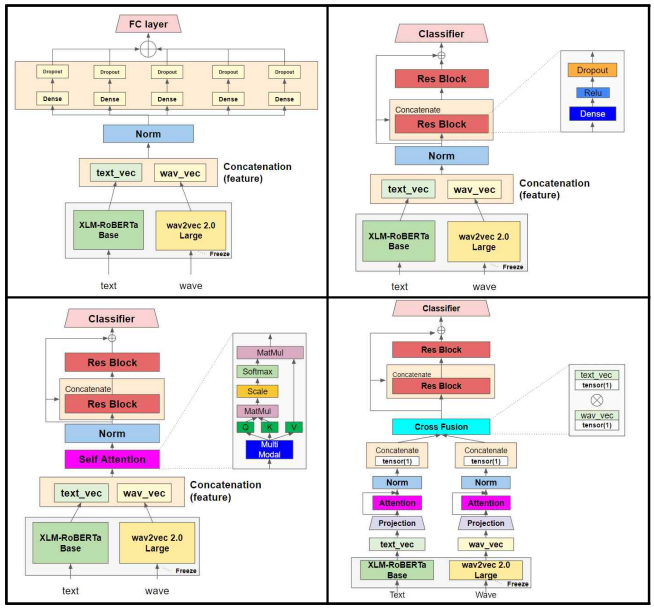

그림 2 전체 모델 구조

Hybrid Fusion 구조는 모달리티별 특징을 일부는 Early Fusion으로 합치고, 일부는 Late Fusion으로 처리하여 최종 결과를 Soft Voting을 통해 출력한다. 이러한 방법은 각 모달리티의 특징을 고려한 조합으로 모델의 성능을 향상시키기 위해 설계했다.

4. Experiments

4.1 실험 환경

학습데이터의 Batch size는 유니모달은 16, 멀티모달은 6과 4로 학습을 진행했다. Optimizer는 AdamW를 사용했으며, Learning Rate는 wav2vec 2.0은 1e-4, RoBERTa 및 멀티모달의 경우에는 1e-5로 적용했다. 본 실험은 RTX 3090 환경에서 진행했다.

4.2 실험 결과

4.2.1 사전학습 다국어 모델과 한국어 모델 간 비교

다음 표 3에서 RSA-CFN을 기준으로 F1-score를 계산하였을 때 한국어 사전학습 모델(KLUE-RoBERTa)이 다국어 사전학습 모델(XLM-RoBERTa)보다 상대적으로 성능이 우수했으며, 이러한 결과는 RSA-CFN 뿐만 아니라 유니모달 및 다른 멀티모달과 비교했을 시에도 동일하게 나타났다.

표 3 PLM 비교

	RSA-CFN (XLM-RoBERTa)	RSA-CFN (KLUE-RoBERTa)
F1-Score(W)	0.89043	**0.89254**

표 4 모델별 성능 평가 결과

Model	F1-Score (W)	Bin-Count							Fusion	Label Transform
		Neutral	Happy	Angry	Surprise	Sad	Fear	Disgust		
Uni-Modal										
① Wav2Vec 2.0(Wav)	0.87966	2,336	224	0	0	0	0	0	-	True
② KLUE-RoBERTa(Text)	0.87552	2,283	232	10	12	8	10	5	-	True
Multi-Modal										
③ Late Fusion	0.88820	2,328	195	11	6	12	4	4	Late	True
④ Early Fusion	0.89021	2,297	189	22	8	18	17	9	Early	True
Fusion Arch										
⑤ MLP-Mixer	0.88520	2,287	199	28	16	14	8	8	Early	True
⑥ Stack	0.89107	2,307	191	22	5	15	16	4	Early	True
⑦ Residual	0.89149	2,257	233	20	20	12	5	13	Early	True
⑧ RSA	0.89165	2,244	247	30	13	9	13	4	Early	True
⑨ RSA-CFN	0.89043	2,338	186	11	5	14	0	6	Early	False
⑩ RSA-CFN	**0.89254**	2,310	193	17	12	11	2	15	Early	True
Hybrid (③+④+⑦+⑩)										
Hybrid	0.88477	2,224	227	24	50	15	12	8	Cross Modality	True

4.2.2 Architecture Fusion 방법론 비교

다음 표 4에서 음성과 텍스트 데이터 기반 Fusion 모델들을 비교한 실험 결과를 보여준다.

유니모달의 경우 텍스트 모델(②)보다 음성 모델(①)의 성능이 87.97%로 더 높았으며 멀티모달을 사용한 Early Fusion(④), Late Fusion(③)이 각각 89.02%, 88.82%로 유니모달 보다 더 높은 성능을 보였다. 이를 통해 유니모달보다 멀티모달을 활용한 Architecture Fusion 모델 성능이 우수한 것을 보여줬다.

또한, Early Fusion 방식에 기반한 다양한 Architecture Fusion 모델을 비교해 본 결과 Label Transform을 적용한 RSA-CFN(⑩)의 성능이 89.25%로 다른 모델들에 비해 상대적으로 제일 우수한 성능을 보여줬다. 이를 통해 RSA-CFN의 멀티모달 간 상호작용을 명시하는 Cross Fusion Layer와 잔차 연결(Residual/Skip Connection) 방법론이 유의미하다는 것을 검증했다.

마지막으로 기존 모델 중 성능이 우수했던 모델들(③+④+⑦+⑩)의 결과를 Soft Voting 한 Hybrid 모델은 다른 모델들에 비해 상대적으로 성능이 낮지만 감정 분류에 대한 다양성이 뛰어난 것을 확인했다.

4.2.3 멀티모달 상호작용 시각화

다음 그림 3에서 음성과 텍스트 벡터 간 상호작용과 중요도를 살펴보기 위해 Self-Attention Weight를 시각화했다. 해당 그림은 Wav와 Text의 상호작용을 살펴보기 위해 검증데이터로 RSA 모델의 Self-Attention Weight를 추출한 후 중립 및 중립 외 라벨별로 평균을 취했고 Vector 별 대푯값으로써 중앙값을 사용했다.

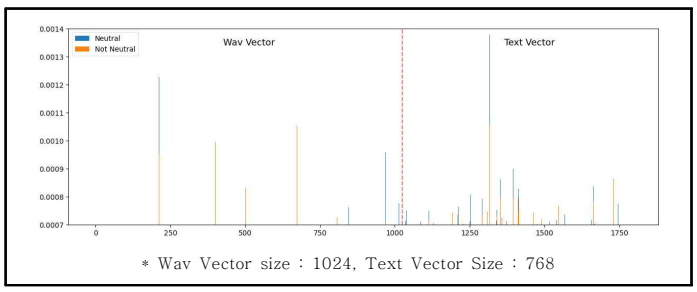

그림 3 Self-Attention Weight

이를 통해 중립인 경우에는 Text Vector가 더 활성화되고 중립이 아닌 감정이 존재하는 경우에는 Wav Vector가 더 활성화됨을 볼 수 있다. 이는 감정이 생기면 사람의 목소리 즉 Wav 데이터에 감정이 실리는 것을 확인할 수 있다.

5. 결론

본 논문에서 제안하고 있는 RSA-CFN은 모달리티별 벡터에 대한 Fusion으로 외적(Cross Product)과 잔차 연결(Skip Connection)을 적용함으로써 각 모달리티 간의 상호작용을 극대화하고 정보손실을 방지하는 것을 유도했다. 이를 통해 단순 결합하는 Architecture Fusion들보다 상대적으로 높은 성능을 검증했다.

하지만, 해당 방법은 많은 리소스가 필요하므로 향후 연구에서는 아키텍처를 효율적으로 설계하여 더 경량화된 모델을 개발해 유사한 성능을 낼 수 있을 것으로 기대한다.

6. 참고문헌

[1] 방나모, 연희연, 이지현, 구명완.(2022).MLP-Mixer 구조를 활용한 대화에서의 멀티모달 감정 인식.한국정보과학회 학술발표논문집,(),2288-2290.

[2] K. J. Noh and H. Jeong, &KEMDy20,& https://nanum.etri.re.kr/share/kjnoh/KEMDy20?lang=ko_KR

[3] Wei, Xi, et al. "Multi-modality cross attention network for image and sentence matching." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. (2020).

[4] Nagrani, Arsha, et al. "Attention bottlenecks for multimodal fusion." Advances in Neural Information Processing Systems 34 (2021): 14200-14213.

[5] 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.

[6] Liu, Yinhan, et al. "Roberta: A robustly optimized bert pretraining approach." arXiv preprint arXiv:1907.11692 (2019).