ISSN 2383-630X(Print) / ISSN 2383-6296(Online)
Journal of KIISE, Vol. 48, No. 10, pp. 1122-1131, 2021. 10
https://doi.org/10.5626/JOK.2021.48.10.1122

SMERT: 감성 분석 및 감정 탐지를 위한 단일 입출력 멀티 모달 BERT

(SMERT: Single-stream Multimodal BERT for Sentiment Analysis and Emotion Detection)

김 경 훈 † 박 진 욱 † 이 지 은 † 박 상 현 ††
(Kyeonghun Kim) (Jinuk Park) (Jieun Lee) (Sanghyun Park)

요 약 감성 분석은 텍스트로부터 주관적인 의견 및 성향을 분석하고, 감정 탐지는 '행복', '슬픔'과 같이 텍스트에서 나타나는 감정을 검출하는 연구다. 멀티 모달 데이터는 텍스트뿐만 아니라 이미지, 음성 데이터가 함께 나타나는 것을 의미한다. 관련 선행 연구에서 순환 신경망 모형 혹은 교차 트랜스포머를 사용한다. 하지만 순환 신경망 모형은 장기 의존성 문제를 가지며, 교차 트랜스포머는 모달리티별 특성을 반영하지 못하는 문제점이 있다. 이를 해결하기 위해 본 연구에서는 멀티 모달 데이터가 하나의 네트워크로 학습되는 단일 입출력 트랜스포머 기반 모형 SMERT를 제안한다. SMERT는 모달리티 결합 표현형을 얻어 이를 감성 분석 및 감정 탐지에 활용한다. 또한, BERT의 훈련 태스크를 멀티 모달 데이터에 활용하기 위해 개량하여 사용한다. 제안하는 모델의 검증을 위해 CMU-MOSEI 데이터셋과 여러 평가 지표를 이용하고, 모달리티 조합별 비교실험과 예시를 통해 모델의 우수성을 검증하였다.

키워드: 자연어 처리, 멀티 모달, 감성 분석, 감정 탐지, 단일 입출력 트랜스포머, BERT

Abstract Sentiment Analysis is defined as a task that analyzes subjective opinion or propensity and, Emotion Detection is the task that finds emotions such as 'happy' or 'sad' from text data. Multimodal data refers to the appearance of image and voice data in addition to text data. In prior research, RNN or cross-transformer models were used, however, RNN models have long-term dependency problems. Also, since cross-transformer models could not capture the attribute of modalities, they got worse results. To solve those problems, we propose SMERT based on a single-stream transformer ran on a single network. SMERT can get joint representation for Sentiment Analysis and Emotion Detection. Besides, we use BERT tasks which are improved to utilize for multimodal data. To present the proposed model, we verify the superiority of SMERT through a comparative experiment on the combination of modalities using the CMU-MOSEI dataset and various evaluation metrics.

Keywords: Natural Language Processing, multimodal, Sentiment Analysis, Emotion Detection, Single-stream Transformer, BERT

· 이 논문은 2021년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원(IITP-2017-0-00477, (SW스타랩) IoT 환경을 위한 고성능 플래시 메모리 스토리지 기반 인메모리 분산 DBMS 연구개발)과 국토교통부의 스마트시티 혁신인재육성사업의 지원을 받아 수행된 연구임

† 학생회원 : 연세대학교 컴퓨터과학과 석사과정
kkh115505@yonsei.ac.kr
parkju536@yonsei.ac.kr
jieun199624@yonsei.ac.kr
†† 종신회원 : 연세대학교 컴퓨터과학과 교수(Yonsei Univ.)
sanghyun@yonsei.ac.kr
(Corresponding author임)

논문접수 : 2021년 4월 20일
(Received 20 April 2021)
논문수정 : 2021년 7월 12일
(Revised 12 July 2021)
심사완료 : 2021년 7월 25일
(Accepted 25 July 2021)

Copyright©2021 한국정보과학회 : 개인 목적이나 교육 목적인 경우, 이 저작물의 전체 또는 일부에 대한 복사본 혹은 디지털 사본의 제작을 허가합니다. 이 때, 사본은 상업적 수단으로 사용할 수 없으며 첫 페이지에 본 문구와 출처를 반드시 명시해야 합니다. 이 외의 목적으로 복제, 배포, 출판, 전송 등 모든 유형의 사용행위를 하는 경우에 대하여는 사전에 허가를 얻고 비용을 지불해야 합니다.
정보과학회논문지 제48권 제10호(2021. 10)

1. 서 론

감성 분석(Sentiment Analysis)은 텍스트로부터 주관적인 태도, 의견 및 성향을 분석하는 자연어 처리 연구다. 그리고 감정 탐지(Emotion Detection)는 '행복'이나 '슬픔'과 같이 텍스트에서 나타나는 감정을 검출하는 연구다. 감성 분석과 감정 탐지는 고객의 의사와 감정을 파악해서 더 나은 서비스를 제공해야 하는 산업 분야에서 활용 가능하다[1].

최근에는 페이스북(Facebook), 인스타그램(Instagram), 유튜브(YouTube)와 같은 소셜 미디어(Social Media)가 성장함에 따라 멀티 모달(Multimodal) 데이터가 많이 생성되고 있다[2-4]. 멀티 모달 데이터란 텍스트 정보뿐만 아니라 이미지, 음성 정보가 함께 주어지는 복합적인 정보를 의미한다. 멀티 모달 데이터 형태로 나타나는 음성 인식, 감성 분석 및 감정 탐지를 인공 지능(AI)을 통해 하려는 연구가 활발히 진행되고 있다.

멀티 모달 연구는 서로 다른 모달리티의 정보를 효과적으로 융합해 정보의 손실을 최소화하는 방향으로 진행되었다. 이는 다음과 같이 크게 세 방식으로 분류할 수 있다. 첫 번째는 데이터의 입력 단계나 예측 단계에서 모달리티를 융합하는 방법이다. 하지만 모달 내(Intra-modal) 및 모달 간(Inter-modal)의 상호 연관성을 잃어버린다는 문제점이 있었다[2]. 두 번째는 결합 표현형(Joint Representation)을 학습하는 방법이다. TFN (Tensor Fusion Network)[2]이나 순환 번역(Cyclic Translation)을 사용하는 계층적 구조를 사용해 결합 표현형을 학습한다[5]. 이러한 방법은 순환 신경망(Recurrent Neural Network) 기반의 모델을 사용하기 때문에 장기 의존성(Long-term dependency) 문제가 발생했다. 마지막은 최근 자연어 처리 연구 분야에서 장기 의존성을 해결하기 위해 주의 기제를 사용하는 트랜스포머(Transformer)[6]를 사용하는 방법이다. 특히 각 모달리티를 입력으로 사용하는 트랜스포머를 서로 교차해 사용하는 방법[3,4,7]이 연구되어 좋은 성능을 보였다.

세 가지 방법 모두 서로 다른 모달리티 정보를 융합하기 위해 특정한 기술들을 사용해야 했고, 이 과정에서 정보의 손실이 발생하게 되었다. 앞선 방법 중 교차 트랜스포머(Cross-transformer) 방법이 정보의 손실을 최소화하여 좋은 성능을 보였으나, 모달리티가 가지고 있는 특징을 반영을 하지 못한다는 단점을 가지고 있다. 대표적인 예시로 감성 분석 및 감정 탐지 연구에서는 텍스트가 결과에 가장 큰 기여를 한다[5]. 하지만 교차 트랜스포머 방법은 모달리티 간의 상호 연관성을 서로 동일한 비율로 적용하기 때문에 상대적으로 다른 방법들보다 텍스트의 정보를 덜 사용하게 되어, 결과에 악영향을 끼치는 경우도 있다[8]. 조건 게이트 메커니즘(Conditional Gated Mechanism)[3]을 사용해서 매 문장마다 반영되는 비율을 조정하는 연구도 있었지만, 이전의 방법들보다 더 많은 계산량을 요구하는 문제점이 있다.

본 논문에서는 앞서 제시한 문제점들을 해결하기 위해 SMERT를 제안한다. SMERT는 모든 모달리티가 하나의 네트워크를 통해 학습되는 단일 입출력 트랜스포머(Single-stream Transformer)를 사용해서 감성 분석 및 감정 탐지를 수행한다. 단일 입출력 트랜스포머는 시각 질의응답(Visual Question & Answering), 이미지 검색(Image Retrieval), 텍스트 검색(Text Retrieval)과 같은 다른 멀티 모달 분야에서 교차 트랜스포머에서 발생하는 문제들을 해결하기 위해 고안된 방법[9-11]이다. 타 연구 분야에서는 이 방법론으로 두 개의 모달리티만을 사용했지만, 본 논문에서는 텍스트, 이미지, 음성 세 개의 모달리티를 하나의 네트워크의 입력으로 사용한다. 또한, 최근 자연어 처리 연구 분야에서 활발하게 사용되고 있는 BERT[12]의 방법인 MLM(Masked Language Modeling)과 NSP(Next Sentence Prediction)를 멀티 모달 데이터에 활용할 수 있도록 수정한 MMLM(Multi-modal masked Language Modeling)과 AP(Alignment Prediction)를 이용해 모달리티별 특성을 반영하는 결합 표현형을 학습하도록 한다.

본 논문의 기여는 다음과 같다. 1) 두 개의 모달리티를 활용했던 이전 연구들과 달리 세 개의 모달리티를 하나의 네트워크로 학습시키는 단일 입출력 트랜스포머 모델을 사용한다. 2) 본 연구에서 BERT의 훈련 태스크에서 착안한 MMLM과 AP를 사용해 모달리티별 특성을 반영하는 결합 표현형을 학습한다. 3) 감성 분석 및 감정 탐지 선행 연구와의 비교실험을 통해 본 모델이 가장 좋은 성능을 가지는 것을 보였다.

본 논문의 구성은 다음과 같다. 2장에서 관련 연구를 소개하고, 3장에서는 본 연구의 문제를 정의한다. 4장에서는 제안하는 모형의 방법론을 제시하고, 5장에서는 데이터셋 소개, 실험 세부사항 및 평가지표에 대해서 기술한다. 6장에서 선행 연구와의 비교 및 생성한 예제에 대해 분석하고, 7장에서 결론 및 향후 계획을 기술한다.

2. 관련 연구

멀티 모달 연구는 서로 다른 모달리티 정보를 효과적으로 융합해 정보의 손실을 최소화하는 방향으로 연구가 진행되고 있다. 가장 먼저 데이터의 입력 단계나 예측 단계에서 모달리티들을 융합하는 연구가 이루어졌다. Early-Fusion[13]은 입력 단계에서 서로 다른 모달리티의 특징들을 합치는 방법이다. 하지만 데이터 처리 단계에서 이루어지는 융합이기 때문에 한 모달리티 내의 상

호작용이 억압되거나 같은 감성 및 감정을 표현하는 모달리티들은 동일한 시간대에 존재해야 한다는 동기화(Synchronized) 문제점을 가지게 된다. Late-Fusion[14]은 각각의 모달리티를 독립된 모델로 구성하고, 이들의 결과물을 다수결 방법(Majority voting)이나 가중 평균(weighted averaging) 방법을 사용해서 결합하는 방법이다. 각 모달리티가 서로 다른 독립적인 네트워크로 존재하기 때문에 모달리티 간의 상호작용을 무시하게 된다. 이러한 문제들을 해결하기 위해서 결합 표현형을 얻는 방법이 연구되었다. Graph-MFN[15]는 그래프 구조와 메모리를 사용하는 방법론을 사용해 결합 표현형을 얻는다. MCTN[5](Multimodal Cyclic Translation Network)은 모달리티 간의 순환 번역을 계층적으로 수행해서 결합 표현형을 얻었다. 하지만 위 모델들은 순환 신경망 계열의 모형을 사용하기 때문에 장기 의존성 문제로 인한 정보의 손실이 발생한다. 최근 자연어 처리 연구 분야에서 장기 의존성 문제를 해결하기 위해 주의 기제 기반의 트랜스포머를 교차로 사용하는 교차 트랜스포머를 이용하여 성능의 발전을 보였다.

일반적으로 모달리티는 독립적으로 가지고 있는 특성이 있다. 교차 트랜스포머를 통해 모든 모달리티가 특성에 관계없이 동일한 비율로 반영된다[3]. 하지만 감성 분석 및 감정 탐지에서는 텍스트의 영향을 가장 많기 때문에 오히려 성능이 좋지 않아지는 경우가 발생하기도 한다. 이를 해결하기 위해서 조건 게이트 메커니즘[3]으로 텍스트, 시각, 음성의 상대적 중요성을 학습하는 방법이 제안되었다. 하지만 다른 방법보다 더 많은 계산량을 요구한다는 단점이 있다.

모든 모달리티가 하나의 네트워크로 학습되는 단일 입출력 트랜스포머는 시각 질의응답, 이미지 검색, 텍스트 검색과 같은 다른 멀티 모달 분야에서 교차 트랜스포머에서 발생하는 문제를 해결하였다. ImageBERT[9]는 이미지와 텍스트를 하나의 입력으로 넣은 후, MLM과 MOC(Masked Object Classification)와 같은 사전 학습 작업(Pre-training task)를 수행하여 이미지 검색, 문장 검색에서 우수한 성능을 보였다. SpeechBERT[10]는 음성과 텍스트를 하나의 입력으로 넣어 음성 질의응답(Speech Question & Answering)에서 뛰어난 성능을 보였다. OSCAR[16]는 객체 태그(Object tag)를 고정점(Anchor points)으로 사용해 표현형을 얻고 여러 분야에 활용 가능하다.

본 연구에서는 텍스트, 이미지, 음성 세 개의 모달리티를 교차 트랜스포머에서 발생하는 문제점을 해결하기 위해 단일 입출력 트랜스포머를 사용하여 모달리티별 특성을 반영하는 결합 표현형을 얻을 수 있다. 또한 학습한 결합 표현형을 감성 분석과 감정 탐지를 위한 분류기에만 사용하였지만, 다른 태스크로도 활용될 수 있는 범용성을 가지고 있다. SMERT 내부에서는 BERT의 MLM과 NSP를 멀티 모달에서 활용할 수 있도록 마스킹(Masking)을 수행하는 MMLM과 올바른 쌍인지 파악하는 AP로 이용해 더욱 성능을 높였다.

3. 문제 정의

본 연구는 텍스트, 이미지, 음성 세 개의 모달리티를 사용하여 감성 분석 및 감정 탐지를 진행한다. 입력 모달리티 X는 다음과 같이 정의한다.

$$X_T \in \mathbb{R}^{d_T \times L}, X_V \in \mathbb{R}^{d_V \times L}, X_S \in \mathbb{R}^{d_S \times L}$$

X_T는 텍스트, X_V는 이미지, X_S는 음성을 의미하며 각각 d_T, d_V, d_S 차원을 가진 L 길이의 시퀀스다. 본 모형은 (X_V, X_S) 쌍을 제외한 나머지 (X_T, X_V), (X_T, X_S)과 같이 모달리티 쌍을 만들어서 사용하고 다음과 같이 정의한다.

$$T = (X_T)$$
$$V' = (X_T, X_V)$$
$$S' = (X_T, X_S)$$

T, V', S'는 한 번의 훈련 스텝에서 순차적으로 포워딩(Forwarding)되고, 역전파(Backpropagation) 과정을 거친다.

멀티 모달 데이터는 “비정렬성(Unalign)”이라는 특성[7]을 가지고 있다. 즉, 각 모달리티가 서로 다른 샘플링 비율(Sampling rate)을 가지고 있어서 데이터가 서로 정확하게 같은 시간대정보를 내포하지 않을 수 있다. 따라서 학습을 진행하기에 앞서 데이터 간의 시간대를 맞추기 위해 정렬(Align)하는 과정이 필요하다. 제안하는 모형에서는 세 모달리티가 같은 길이를 가질 수 있도록 피벗(Pivot) 모달리티를 텍스트로 정하고, 컬랩스 함수(Collapse)로 평균 함수를 사용해 정렬한다.

감성(Y_S)과 감정(Y_E)은 텍스트에서 나타나는 라벨을 기준으로 하며, Y_S는 [-3, +3] 범위의 실수이며, Y_E는 0과 1로 분류된다.

4. 모델 구조

본 연구에서 제안한 SMERT 모델의 전체적인 구조는 Fig. 1과 같다. 제안하는 모델은 다음과 같이 크게 세 부분으로 구분할 수 있다. 첫 부분은 결합 임베딩으로 융합 게이트(Fusion Gate)를 사용하여 모달리티들을 융합한다. 두 번째 부분인 SMERT는 텍스트, 이미지, 음성에 대해 단일 입출력 트랜스포머 모델을 사용하고, BERT에서 사용하는 MLM과 NSP 작업을 멀티 모달 데이터에 사용할 수 있도록 수정한 MMLM과 AP를 사용한다. 마지막으로 감성 분석 및 감정 탐지를 위한 분류기를 사용한다.

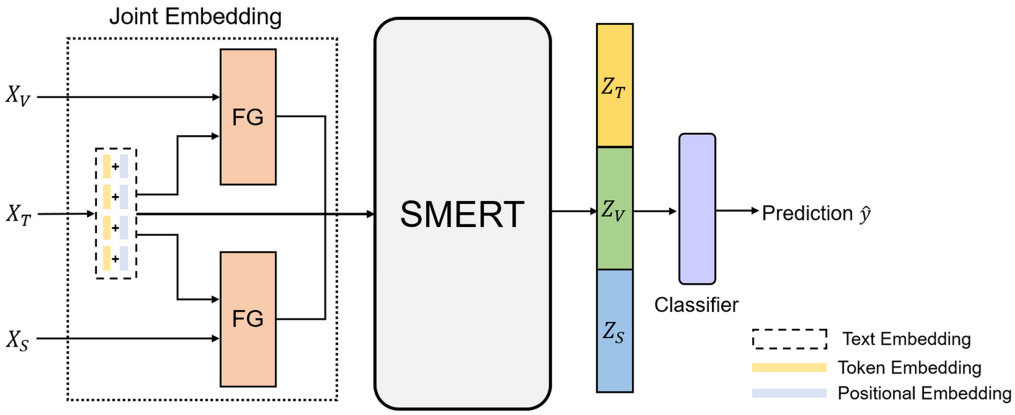

그림 1 제안하는 모형인 SMERT의 전체적인 흐름도
Fig. 1 A Flow Chart of the proposed method, SMERT

4.1 결합 임베딩(Joint Embedding)

Fig. 1에서 점선에 해당하는 부분인 결합 임베딩은 텍스트 임베딩(Text Embedding)과 융합 게이트로 구성된다. 텍스트 X_T는 단어들의 집합이므로 토큰 임베딩(Token Embedding)과 위치 임베딩(Positional Embedding)이 있는 텍스트 임베딩을 통과한다. 이는 토큰 임베딩을 통해 단어를 실수 값으로 변환하고, 연속적인 특성을 가지고 있는 텍스트에 위치 정보를 부여하기 위함이다. 토큰 임베딩과 위치 임베딩은 BERT에서 사용한 것을 차용한다. X_V, X_S와 같은 경우에는 연속적인 위치 정보를 가지고 있지 않기 때문에 위치 임베딩을 수행하지 않는다. 텍스트 임베딩을 통해 얻은 값은 융합 게이트의 입력으로 사용되며, 융합 게이트는 다음과 같다.

4.1.1 융합 게이트

SMERT에는 T, V', S'와 같이 세 가지 쌍을 만들어서 입력으로 사용한다. 이 중 V', S'는 융합 게이트에서 쌍이 이루어지는 작업을 수행한다. 융합 게이트는 Fig. 2와 같이 구성되어 있다. 두 개의 모달리티를 결합하고, 이를 구분하기 위한 분할 임베딩(Segment Embedding) 값을 더한다. 마지막으로 텍스트가 아닌 모달리티의 특성으로 증폭시키기 위해 해당 값을 곱한 뒤 LayerNorm을 수행한다. 수식은 다음과 같다.

$$A \oplus B = B' \cdot ([A; B'] + Segment) \quad (1)$$

$$B' = Linear(B) \quad (2)$$

$$X_{V'} = \mathrm{LN}(X_T \oplus X_V),\ X_{S'} = \mathrm{LN}(X_T \oplus X_S) \quad (3)$$

$A \oplus B$는 'A와 B를 융합한다'라고 정의하며 A는 텍스트를 B는 텍스트를 제외한 다른 모달리티를 의미한다. 따라서 B를 결합하고 곱하는 과정을 위해 Linear를 통해 텍스트 차원인 d_T로 투영한다. $[A; B']$에서 A와 B'를 시퀀스의 길이를 기준으로 결합하고, 분할 임베딩 값을 더한다. 그리고 $\cdot$를 통해 B'의 값을 모달리티에 각각 곱해 텍스트가 아닌 모달리티의 특성으로 증폭시킨다. 마지막으로 d_T의 정규화 차원을 가지고 있는 LN은 LayerNorm을 의미하며 $X_{V'}$과 $X_{S'}$는 $X_{V'}, X_{S'} \in \mathbb{R}^{dT \times 2L}$와 같은 차원과 길이를 가진다.

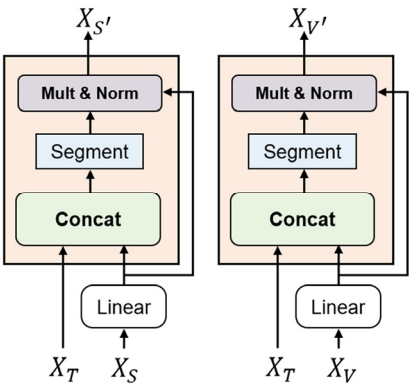

그림 2 융합 게이트
Fig. 2 The Fusion Gate

4.2 SMERT

SMERT는 Fig. 3과 같이 동작한다. 결합 임베딩으로 얻은 값들은 E 토큰들로 표현되고, SMERT를 통해 나온 결과물은 H 토큰들로 표현한다. H 토큰들 중에 $H_{<CLS>}$ 토큰은 다운스트림(Downstream) 태스크에 사용할 수 있는 Pooler에 해당한다. 따라서 이 Pooler에 완전 연결 신경망(Fully Connected Layer)과 하이퍼탄젠트(tanh)를 사용한다. T, V', S' 세 쌍에 대해 각각 수행하며 수식 (3)과 같은 과정으로 Fig. 1의 Z_T, Z_V, Z_S를 얻는다.

$$Z_{\{T,V,S\}} = \tanh\left(Linear(H_{<CLS>})\right) \quad (4)$$

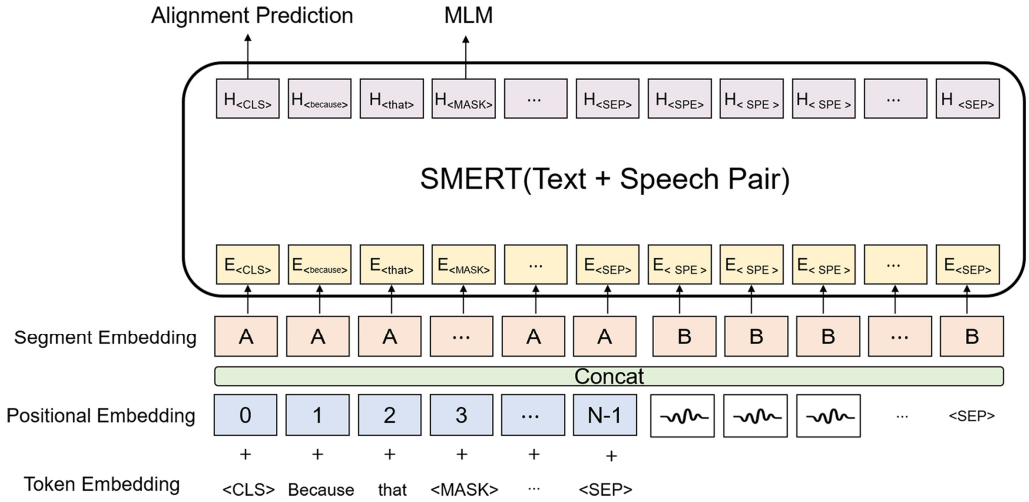

그림 3 텍스트, 음성일 때, SBERT 내부 구조

Fig. 3 Internal Structure of SBERT when it gets text and speech

4.2.1 MMLM (Multimodal Masked Language Modeling)

이 태스크는 BERT의 MLM 태스크와 유사하다. T, V', S' 세 쌍에서 모두 텍스트에 대해서만 마스킹을 수행한다. 각 훈련 텍스트 샘플에 대해서 토큰 위치의 15%를 임의로 선택한다. 만약 i 번째 토큰이 선택되면, i 번째 토큰은 80%의 확률로 [MASK]되고, 10%의 확률로 랜덤한 토큰으로 변경되고, 나머지 10%의 확률로 변동되지 않고 그대로 남는다.

4.2.2 AP(Alignment Prediction)

이 태스크는 BERT의 NSP 태스크와 유사하다. V', S' 두 쌍에서만 진행하며, 두 모달리티 간의 관계성을 파악하기 위해서 진행한다. 예를 들면, V' 쌍에서, 훈련 샘플에서 X_T와 X_V는 다음과 같이 선택한다. X_V는 80%의 확률로 실제 X_T와 쌍을 이루는 문장이다(IsPair). X_V는 20%의 확률로 X_T와 관계없이 임의로 선택된 문장이다(UnPair). 쌍을 이루는 문장들의 IsPair와 UnPair은 $y_p \in \{0,1\}$로 표시한다.

4.3 분류기

각 T, V', S' 쌍의 $H_{<CLS>}$로부터 얻은 Z_T, Z_V, Z_S 값들의 정보를 결합하고, 이를 감성 분석 및 감정 탐지에 활용하기 위해 GRU(Gated Recurrent Unit) 신경망을 사용한다.

$$P_y = GRU(Z_T, Z_V, Z_S) \tag{5}$$

최종적으로 분류를 위해 P_y를 완전 연결 신경망에 통과한다. 이 때, 감성 분석에서는 하이퍼탄제트를 사용하고, 감정 탐지에서는 소프트맥스(softmax)를 사용한다.

$$\hat{Y}_S = \tanh\left(Linear(P_y)\right) \tag{6}$$

$$\hat{Y}_E = softmax(Linear(P_y)) \tag{7}$$

4.4 손실 함수

감성 분석과 감정 탐지는 서로 다른 손실 함수를 사용한다. 감성 분석의 경우에는 SMERT에서 발생하는 MMLM과 AP의 손실 함수와 분류기에서 발생하는 감성 분석 예측 손실 함수로 구성된다. MMLM과 AP는 각각 CrossEntropy 손실 함수를 사용하며, L_{MMLM}과 L_{AP}와 같이 표기한다. T, V', S' 세 쌍에 대해 각각 수행하고 합한 것을 결합 손실 함수(Joint loss)라 지칭하며, L_j와 같이 표기한다.

$$L_J = L_{MMLM} + L_{AP},\ J \in \{T, V', S'\} \tag{8}$$

L_j는 T, V', S'에서 각각 이뤄지므로 L_T, $L_{V'}$, $L_{S'}$와 같이 세 개가 존재한다. 감성 분석 손실 함수 $L_{Sentiment}$은 MSE 손실 함수와 L_T, $L_{V'}$, $L_{S'}$의 평균 값을 더해 구한다. 수식은 다음과 같다.

$$L_{Sentiment} = L_{MSE} + \frac{(L_T + L_{V'} + L_{S'})}{3} \tag{9}$$

감정 탐지의 경우에도 동일하게 결합 손실 함수 L_j와 감정 탐지 예측 손실 함수를 사용한다. 감정 탐지 손실 함수은 CrossEntropy 손실 함수와 L_T, $L_{V'}$, $L_{S'}$의 평균 값을 더해 구한다. 수식은 다음과 같다.

$$L_{Emotion} = L_{CE} + \frac{(L_T + L_{V'} + L_{S'})}{3} \tag{10}$$

5. 실험 세부사항

5.1 데이터셋

본 연구는 멀티 모달 감성 분석 및 감정 탐지를 할 수 있는 CMU-MOSEI(CMU Multimodal Opinion Sentiment and Emotion Intensity) 데이터셋[15]을 사용해

실험을 진행한다. CMU-MOSEI 데이터셋은 총 65시간의 주석이 달린 유튜브의 영화 리뷰 클립에서 1000명의 서로 다른 화자와 250개의 주제로 23,453개의 문장을 추출해 감성과 6개의 감정을 라벨링(labeling)했다. 감성은 CMU-MOSI(CMU-Multimodal Opinion-level Sentiment Intensity) [17]와 스탠포드 감성 트리뱅크(Stanford Sentiment Treebank)[18]의 주석을 다는 방식에 따라 [-3,3] 범위의 점수로 스케일링(Scaling)되었다. 각 감성 점수가 의미하는 것은 다음과 같다. [-3: 강한 부정(HN), -2: 부정(Negative), -1: 약한 부정(WN), 0: 중립(Neutral), +1: 약한 긍정(WP), +2: 긍정(Positive), +3: 강한 긍정(HP)]. 해당 점수들의 분포는 Fig. 4의 위 그래프와 같다. 감정은 에크만 감정(Ekman Emotions) [19]에 따라 {행복, 슬픔, 화남, 공포, 역겨움, 놀람}으로 구성되어 있으며, [0,3]의 값으로 스케일링 되어 있다. 각 감정 점수가 의미하는 것은 다음과 같다. [0: 감정이 나타나지 않음(Non-Emotion), 1: 약한 감정(WE), 2: 감정(Emotion), 3: 강한 감정(HE)]. 감정들의 분포는 Fig. 4의 아래 그래프와 같다. 감정은 한 문장에서 동시에 여러 감정이 나타날 수 있는 멀티 감정 라벨(Multi-emotion label)이다. 훈련 데이터셋의 크기는 16,315개, 검증 데이터셋의 크기는 1,871개, 테스트 데이터셋의 크기는 4,654개로 관련 연구 데이터셋으로 사용되는 CMU-Multimodal SDK[20]에서 정해진 것을 사용한다.

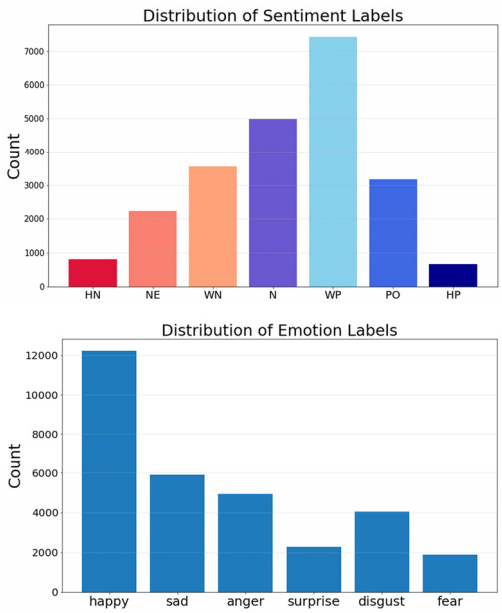

그림 4 감성 라벨의 분포 및 감정 라벨의 분포

Fig. 4 Distribution of Sentiment Labels and Distribution of Emotion Labels

5.2 평가 지표

본 연구에서 감성 분석은 A7정확도, A2정확도, MAE (Mean Absolute Error), F1 score를, 감정 탐지는 A2 정확도, F1 score를 이용한다.

A7정확도는 양수면 올림, 음수면 내림을 하는 스케일링 을 통해 [-3,3]의 라벨을 맞추도록 하는 평가방법이다. A2정확도는 감성 분석에서는 $Y_s \geq 0$ 이면 긍정 감성, $Y_s \leq 0$ 이면 부정 감성으로 구분하는 이진 분류를 한다. 감정 탐지에서는 $Y_E > 0$ 이면 감정 표출, $Y_E = 0$ 이면 감정 없음을 구분하는 이진 분류를 한다.

MAE는 평균절대오차로 모든 절대 오차의 평균을 의미한다. 다음과 같은 수식으로 계산한다.

$$MAE = \frac{1}{n}\sum\left|Y - \hat{Y}\right| \tag{11}$$

F1 score의 경우에는 Precision과 Recall의 조화 평균이다. Precision은 참이라고 예측한 비율 중 진짜 참의 비율이고, Recall은 실제 참 데이터 중 참이라고 예측한 비율이다. 다음과 같은 공식으로 계산한다.

$$F = 2 \times \frac{precision \times recall}{precision + recall} \tag{12}$$

5.3 훈련 세부사항

각 모달리티의 특징 추출은 CMU-Multimodal SDK를 사용한다. 텍스트 특징 추출은 BERTTokenzier를 사용하였으며, 768차원(d_t)을 가지고 있다. 이미지 특징 추출은 비디오 클립에서 화자의 얼굴 표정에 대한 특징을 추출해야 하기 때문에, 얼굴 표정 분석 툴킷인 FACET[21]을 사용했으며, 35차원(d_v)을 가지고 있다. 음성 특징 추출기로 COVAREP[22]을 사용하였으며, 74차원(d_s)을 가지고 있다. 이 세 모달리티의 차원은 결합 임베딩을 거친 후에 768차원으로 통일된다. 시퀀스 길이 L은 100으로 설정한다. 정규화(Regularization)을 위한 드롭 아웃(Dropout) 비율을 0.5로, 가중치 감쇠(Weight Decay) 값은 0.01로 부여하였으며, 학습률은 0.0001, 배치 사이즈(batch size)는 8로 설정하였다.

6. 실험 결과 및 토의

6.1 선행연구와의 비교

감성 분석에 대한 실험 결과는 Table 1과 같다. SMERT는 다른 모형들과 비교하여 A2 정확도, A7 정확도, MAE, F1 score 모든 평가 지표 부문에서 가장 좋은 성능을 기록하였다. 이는 제안하는 모델이 선행연구보다 성능이 우수함을 보여준다.

정확도 관련 평가지표에서 LF-LSTM와 같은 Fusion 계열보다 A2 정확도는 3.7점이 A7 정확도는 18점이 높으며, MCTN과 같은 순환 신경망 계열의 모델들 보다 A2 정확도는 4.5점이 A7 정확도는 16.8점이 높다. 순환

표 1 선행 연구와의 감성 분석 성능 비교
Table 1 The performance comparison with previously studied models about Sentiment Analysis

Metric	A2 (↑)	A7(↑)	MAE(↓)	F1(↑)
EF-LSTM[7]	78.2	47.4	0.642	77.9
LF-LSTM[7]	80.6	48.8	0.619	80.6
Graph-MFN[15]	76.9	45.0	0.710	77.0
RAVEN[8]	79.1	50.0	0.614	79.5
MCTN[5]	79.8	49.6	0.609	80.6
MulT[7]	82.5	51.8	0.580	82.3
Self-MM[23]	82.8	-	0.530	82.5
Multilogue-net[4]	82.1	-	0.590	80.0
SMERT	**84.3**	**66.8**	**0.344**	**84.0**

신경망 기반 모델인 EF-LSTM과 LF-LSTM보다 트랜스포머 기반인 제안하는 모델 SMERT가 모든 평가지표에서 더 나은 성능을 보이기 때문에, 장기 의존성 문제를 해결한 것을 보여준다[6]. 또한 모달리티 간의 상호 연관성을 전혀 고려하지 않는 모델인 LF-LSTM과 모달리티 간의 상호 연관성만을 고려하는 EF-LSTM보다 SMERT가 더 뛰어난 성능을 보이므로, 선행연구보다 모달리티 간의 상호 연관성을 잘 고려하여 결과를 예측하는 것을 알 수 있다. 교차 트랜스포머 모델인 MulT에 비해서 A2 정확도는 1.8점, A7 정확도는 15점이 높아 단일 입출력 트랜스포머의 우수성을 알 수 있다. 특히 A7정확도과 MAE는 기존의 모델들보다 더 많은 점수의 향상을 보였는데, 제안하는 모델이 실제 라벨을 매우 유사하게 예측하고 있음을 보여준다. 따라서 제안하는 방법론이 정교한 예측이나 세분화된 라벨 예측에서 더욱 강점을 보이는 것을 알 수 있다.

감정 탐지에 대한 실험 결과는 Table 2와 같으며, A2 정확도와 F1 score로 성능을 비교한다. 감정 탐지는 같은 모형을 사용해 각 감정을 검출하는 방식으로 실험을 진행했다. G-MFN(Graph-MFN)에서 A2 정확도는 가중된(weighted) 방법을, TBJE[24]와 SMERT에서는 표준 방법을 사용했다. 비교의 정확성을 위해 G-MFN과는 A2 정확도를 비교하지 않았다. TBJE의 경우 TBJE-2와 TBJE-3이 있는데 2와 3은 사용한 모달리티를 의미하며, TBJE-2가 성능이 더 좋지만, 실험의 공정성을 위해서 TBJE-3을 같이 서술한다.

'Sad', 'Surprise'를 제외하고, 나머지 네 개 감정에서 A2 정확도 성능의 향상을 보였으며 특히 'Happy'의 경우 2.5점만큼 성능이 높아졌다. 그리고 'Sad'나 'Surprise'도 선행 연구와 근접한 성능을 보인다. F1 score은 'Fear'와 'Surprise'를 제외하고 성능의 향상을 보였다. 각 감정에 대한 성능을 보았을 때, 'Happy'만이 높은 점수를 가지고 있지 않다. 이는 Fig. 4의 감정 라벨의 분포를 보았을 때 'Happy'를 제외한 다른 감정들의 개수들이 최대 6,000개 정도만 존재하고, 총 데이터 개수인 23,453개에 비해 적게 나타나는 것을 알 수 있다. 즉, 데이터가 불균형하게 구성되어 있고, 해당 감정들은 점수가 높게 나타난다.

6.2 모달리티 조합에 따른 비교실험

세 개의 모달리티를 모두 활용하는 것이 단일로 사용하는 것보다 우수하다는 점을 검증하기 위하여 모달리티를 하나씩 제외하여 비교 실험을 진행한다. 실험은 감성 분석으로 진행을 하였으며, Table 3은 세 개의 모달리티를 사용한 경우인 $T+V'+S'$, 두 개인 $T+V'$, $T+S'$, $V'+S'$, 하나씩 사용한 T, V', S'를 A2 정확도, MAE, F1 score 로 성능을 비교한 결과다.

단일 모달리티의 결과들을 보면 경우에 T가 V', S'보다 A2가 4점 이상 더 높은 것을 알 수 있다. 이를 통해 텍스트, 이미지, 음성 세 개의 모달리티 중에서 텍스트가 감성 분석을 진행하는데, 가장 큰 영향을 가지는 것을 알 수 있다. 하나 이상의 모달리티를 사용한 경우를 보면 텍스트 하나의 경우를 사용한 것보다 추가적으로

표 2 선행 연구와의 감정 분류 성능 비교
Table 2 The performance comparison with previously studied models about Emotion Detection

Emotion	Happy		Sad		Anger		Fear		Disgust		Surprise	
Metric	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
G-MFN[15]	-	66.3	-	66.9	-	72.8	-	**89.9**	-	76.6	-	85.5
TBJE-2[24]	66.0	65.5	**73.9**	67.9	81.9	76.0	89.2	87.2	86.5	84.5	**90.6**	**86.1**
TBJE-3[24]	65.0	64.0	72.0	67.9	81.6	74.7	89.1	84.0	85.9	83.6	90.5	**86.1**
SMERT	**67.5**	**67.5**	**73.8**	**69.6**	**82.2**	**77.7**	**90.6**	**86.1**	**86.5**	**85.3**	**89.2**	**84.5**

표 3 모달리티 조합 별 성능 비교
Table 3 The performance comparison between modalities combinations

Modality	A2(↑)	MAE(↓)	F1 (↑)
$T+V'+S'$	**84.3**	0.344	**84.0**
$T+V'$	83.4	0.363	83.4
$T+S'$	84.1	0.359	83.9
$V'+S'$	78.4	0.389	77.8
T	83.1	**0.336**	82.9
V'	78.8	0.374	78.5
S'	79.1	0.366	79.1

이미지나 음성을 함께 훈련할 때 성능이 향상되는 것을 보였고, 텍스트, 이미지, 음성을 모두 사용한 경우에는 텍스트만을 사용한 경우보다 A2가 1.2점 더 높은 것을 알 수 있다. 실험의 결과로 보아 세 개의 모달리티를 모두 활용할 때 더 우수하다는 것을 알 수 있다.

6.3 정성적 평가

Table 4는 5.2 모달리티 조합에 따른 비교실험에서 $T+V'+S'$과 T 모델의 감성 분석 예측 결과 예시다. 이미지와 음성은 프레임 단위로 구성되어 있고, 하나의 표로만 표현하기 힘들어 지면상 모두 표시하지 못하고 텍스트 문장과 해당하는 라벨 값들만 설명한다.

첫 번째 문장은 “끔찍한 프롬 중에서도 좋은 순간이 있어서 다행”이라는 뜻을 가지고 있는데, 텍스트에서는 However, disastrous와 같이 부정적인 단어를 통해 ‘부정적’이라고 예측했지만 세 모달리티 모델에서는 얼굴 표정과 음성을 듣고 ‘긍정적’이라고 옳게 예측하는 것을 알 수 있다. 두 번째 문장에서도 텍스트에서는 부정적인 단어만을 보고 ‘부정적’이라고 예측했지만, 세 모달리티를 사용한 경우에는 문장의 의미를 정확히 파악해 정확하게 예측을 진행했다. 세 번째 문장의 경우에는 코미디 영화에 대한 리뷰에 대한 내용이다. “코미디 영화의 내용이 너무 유치하고 재미없어서 아이들이나 볼 것 같은 영화”라고 매우 강조하고 있다. 텍스트만을 사용한 모델에서는 문장에서 별 다른 부정적인 단어가 없고, comedy, kid와 같이 긍정적인 단어만이 나타나고 있기 때문에 ‘긍정적’이라고 예측을 진행했다. 하지만 세 모달리티를 사용한 경우에 얼굴 표정과 비꼬아서 말하는 음성으로 인해 ‘부정적’이라고 올바르게 예측한 것을 알 수 있다. 네 번째 문장에서도 영화가 재미없다는 사실을 직접적으로 말하지 않고, 돌려서 말하고 있는 것을 알 수 있다. 텍스트의 경우 이러한 사실을 알지 못하고 ‘긍정적’이라고 예측했지만 세 모달리티를 사용하면 얼굴 표정과 음성을 통해 올바르게 ‘부정적’이라고 예측했다.

반면에 다섯, 여섯 번째 문장의 경우에는 텍스트만을 사용한 것이 더 정확하게 예측했다. 이 두 문장을 보면 “단지 짧은 영화다.”, “저는 변호사이면서 심리학자인 유일한 대통령 후보다.”와 같이 감성이 잘 나타나지 않는 평서문임을 알 수 있다. 이 예시들에서는 텍스트 이외의 모달리티 정보가 추가적으로 많이 필요하지 않지만 정보를 포함하게 되어 예측을 방해했다. 하지만 대부분의 경우에는 세 모달리티를 모두 사용해서 감성 분석을 예측하는 것이 텍스트 하나보다 더 높은 예측을 하는 것을 알 수 있다.

7. 결 론

본 연구에서는 세 모달리티의 결합 표현형을 얻기 위해, MLM과 AP 태스크를 수행하는 단일 입출력 트랜스포머를 제안한다. 감성 분석 및 감정 탐지 선행 연구들은 순환 신경망 계열의 모델이나 교차 트랜스포머를 사용해서 예측을 한다. 하지만 순환 신경망 모형에선 장기 의존성 문제가 발생하며, 교차 트랜스포머는 모달리티마다 다른 특성을 반영하지 못하는 한계를 보였다. 이러한 한계점을 극복하기 위해 단일 입출력 트랜스포머를 제안하여, 모달리티별 특성을 반영하는 결합 표현형을 얻었다. 제안하는 모델의 성능을 검증하기 위해 CMU-MOSEI 데이터셋으로 감성 분석 및 감정 탐지 실험을 진행하고, A2 정확도, F1 score와 같은 여러 평가 지표를 이용하여 선행 연구보다 성능이 우수함을 보였다. 그리고 예시를 통해 제안한 방법론이 세 모달리티

표 4 세 모달리티과 텍스트 모달리티만을 사용한 모델들의 예측 예시
Table 4 Examples of the prediction from three modalities and text modality

Sentence	Ground Truth	$T+V'+S'$	T
However fortunately this disastrous prom does have its heartfelt moments	**Positive**	**Positive**	Negative
The movie I cannot even stop laughing	**Positive**	**Positive**	Negative
It was supposed to be a comedy an animation and perhaps a kid's movie and it was that it was a kid's	**Negative**	**Negative**	Positive
It is a long movie it's like two and a half hours maybe almost two to forty-five minutes	**Negative**	**Negative**	Positive
It's actually kind of a short movie it's only like	**Negative**	Positive	**Negative**
I am the only candidate for president who is both a lawyer and a psychologist	**Positive**	Negative	**Positive**

를 모두 사용할 때 성능 개선에 도움이 됨을 보여주었다.

향후 연구로는 모달리티를 두 개의 쌍으로 묶어 입력값으로 사용하는 것이 아닌, 세 모달리티를 하나의 쌍으로 만들어 입력값으로 활용해 훈련시간 개선 및 더 나은 결합 표현형을 얻는 연구를 진행할 것이다. 또한, 본 연구에서 사용한 CMU-MOSEI 이외에 감성 분석 및 감정 탐지 연구에서 사용되는 다른 데이터셋인 CMU-MOSI, IEMOCAP[25]에 대해 실험을 수행할 계획이다.

References

[1] A. Shathik and K.P. Karani, "A Literature Review on Application of Sentiment Analysis Using Machine Learning Techniques," *International Journal of Applied Engineering and Management Letters (IJARML)*, Vol. 4, No. 2, pp. 41-77, Aug. 2020.

[2] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.P. Morency, "Tensor fusion network for multimodal sentiment analysis," *arXiv preprint arXiv:1707.07250*, 2017.

[3] A. Kumar and J. Vepa, "Gated Mechanism for Attention Based Multimodal Sentiment Analysis," *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE*, pp. 4477-4481, Feb. 2020.

[4] A. Shenoy and A. Sardana, "Multilogue-net: A context aware rnn for multi-modal emotion detection and sentiment analysis in conversation," *arXiv preprint arXiv:2002.08267*, 2020.

[5] H. Pham, P.P Liang, T. Manzini, L.P Morency, and B. Póczos, "Found in translation: Learning robust joint representations by cyclic translations between modalities," *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, No. 1, pp. 6892-6899, Jul. 2019.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Usz-koreit, L. Jones, A.N. Gomez, Ł. Kaiser, and I. Po-losukhin, "Attention is all you need", *Advances in neural information processing systems*, pp. 5998-6008, 2017.

[7] Y.H.H Tsai, S. Bai, P.P. Liang, J.Z. Kolter, L.P. Morency, and R. Salakhutdinov, "Multimodal Transformer for Unaligned Multimodal Language Sequences," *Proceedings of the conference. Association for Computational Linguistics. Meeting*, Vol. 2019, pp. 6558-6569, Jul. 2019.

[8] Y. Wang, Y. Shen, Z. Liu, P.P. Liang, A. Zadeh, L.P. Morency, "Words Can Shift: Dynamically Adjusting Word Representations Using Nonverbal Behaviors," *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, No. 1, pp. 7216-7223, Jul. 2019.

[9] D. Qi, L. Su, J. Song, E. Cui, T. Bharti, A. Sacheti, "ImageBERT: Cross-modal Pre-training with Large-Scale Weak-Supervised Image-Text Data," *arXiv preprint arXiv:2001.07966*, 2020.

[10] Y.S. Chuang, C.L. Liu, H. Lee, and L. Lee, "SpeechBERT: An Audio-and-text Jointly Learned Language Model for End-to-end Spoken Question Answering," *arXiv preprint arXiv:1910.11559*, 2019.

[11] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, "Vl-bert: Pre-training of generic visual-linguistic representations," *arXiv preprint arXiv:1908.08530*, 2019.

[12] J. Devlin, M.W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[13] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, "Convolutional MKL Based Multimodal Emotion Recognition and Sentiment Analysis," *2016 IEEE 16th international conference on data mining (ICDM), IEEE*, pp. 439-448, 2016.

[14] B. Nojavanasghari, D. Gopinath, J. Koushik, T. Baltrušaitis, and L.P. Morency, "Deep Multimodal Fusion for Persuasiveness Prediction," *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pp. 284-288, 2016.

[15] A. Zadeh, P.P. Liang, J. Vanbriesen, S. Poria, E. Tong, E. Cambria, M. Chen, and L.P. Morency, "Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph," *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Vol. 1, pp. 2236-2246, Oct. 2016.

[16] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, Y. Choi, and J. Gao, "Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks," *European Conference on Computer Vision*, pp. 121-137, Aug. 2020.

[17] A. Zadeh, R. Zellers, E. Pincus, and L. P. Morency, "MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos," *arXiv preprint arXiv:1606.06259*, 2016.

[18] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. Manning, A. Ng, and C. Potts, "Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank," *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631-1642, Oct. 2013.

[19] P. Ekman, W.V. Freisen, and S. Ancoli, "Facial Signs of Emotional Experience," *Journal of personality and social psychology*, Vol. 39, No. 6, pp. 1125-1134, 1980.

[20] A. Zadeh, P.P. Liang, S. Poria, P. Vij, E. Cambria, and L.P. Morency, "Multi-attention Recurrent Network for Human Communication Comprehension," *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32, No. 1, Apr. 2018.

[21] T.B Moeslund, A. Hilton, V. Krüger, and L. Sigal, "Visual Analysis of Humans," *Springer*, London, 2011.

[22] G. Degottex, J. Kane, T. Drugman, T. Raition, and S. Scherer, "COVARER - A Collaborative Voice Analysis Repository for Speech Technologies," *2014 ieee international conference on acoustics, speech and signal processing (icassp). IEEE*, pp. 960-964, May. 2014.
[23] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning Modality-specific Representations with Self-Supervised Multi-Task Learning for Multimodal Sentiment Analysis," *arXiv preprint arXiv:2102.04830*, 2021.
[24] J.B. Dellbrouck, N. Tits, M. Brousmiche, and S. Dupont, "A Transformer-based joint-encoding for Emotion Recognition and Sentiment Analysis," *arXiv preprint arXiv:2006.15955*, 2020.
[25] C. Busso, M. Bulut, C.C. Lee, A. Kazemzadeh,E. Mower, S. Kim, J. N. Chamg. S. Lee, and S.S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, Vol. 42. No. 4, pp. 335-359, 2008.

김 경 훈
정보과학회논문지
제 48 권 제 8 호 참조

박 진 욱
정보과학회논문지
제 48 권 제 3 호 참조

이 지 은
정보과학회논문지
제 48 권 제 3 호 참조

박 상 현
정보과학회논문지
제 48 권 제 3 호 참조