영상 데이터 감정 분류를 위한 멀티 모달 기반의 ViT 모델

김예림[O], 이동규[*], 안서영[*], 김지현[*]
[O]성신여자대학교 미래융합기술공학과,
[*]성신여자대학교 미래융합기술공학과
e-mail: {220216126, 220216138}@sungshin.ac.kr[O*], 98_0731@honglab.org[*], jeehyunee3@naver.com[*]

Multi-Modal based ViT Model for Video Data Emotion Classification

Yerim Kim[O], Dong-Gyu Lee[*], Seo-Yeong Ahn[*], Jee-Hyun Kim[*]
[O]Department of Future Convergence Technology Engineering Sungshin Women's University,
[*]Department of Future Convergence Technology Engineering Sungshin Women's University

● 요 약 ●

최근 영상 콘텐츠를 통해 영상물의 메시지뿐 아니라 메시지의 형식을 통해 전달된 감정이 시청하는 사람의 심리 상태에 영향을 주고 있다. 이에 따라, 영상 콘텐츠의 감정을 분류하는 연구가 활발히 진행되고 있고 본 논문에서는 대중적인 영상 스트리밍 플랫폼 중 하나인 유튜브 영상을 7가지의 감정 카테고리로 분류하는 여러 개의 영상 데이터 중 각 영상 데이터에서 오디오와 이미지 데이터를 각각 추출하여 학습에 이용하는 멀티 모달 방식 기반의 영상 감정 분류 모델을 제안한다. 사전 학습된 VGG(Visual Geometry Group)모델과 ViT(Vision Transformer) 모델을 오디오 분류 모델과 이미지 분류 모델에 이용하여 학습하고 본 논문에서 제안하는 병합 방법을 이용하여 병합 후 비교하였다. 본 논문에서는 기존 영상 데이터 감정 분류 방식과 다르게 영상 속에서 화자를 인식하지 않고 감정을 분류하여 최고 48%의 정확도를 얻었다.

키워드: 영상 콘텐츠(Video content), 감정분류(emotion classification), VGG(Visual Geometry Group), ViT(Vision Transformer)

I. Introduction

최근 OTT 서비스(Over-The-Top media service) 의 보급이 확대됨에 따라 영상 콘텐츠의 소비가 증가하였고, 영상 콘텐츠를 통해 영상물의 메시지뿐 아니라 메시지의 형식을 통해 전달된 감정이 시청하는 사람의 감정에 영향을 주고 있다. 이에 따라, 영상 콘텐츠의 감정을 분류하는 연구가 활발히 진행되고 있으며, 특히 영상의 오디오와 이미지를 이용한 멀티 모달 감정인식 기술의 연구가 활발히 진행되고 있다.

한편, NLP(Natural Language Processing) 분야에서 사용하는 Transformer 모델을 image classification 분야에 맞게 변형한 ViT[1]은 기존 Vision task 분야에서 높은 성능을 보이던 CNN(Convolution Neural Network)을 사용하지 않고 여러 데이터 셋(ImageNet, ImageNet-Real, CIFAR-100)에서 SoTA(State of The Art) 성능을 보여주었다.

본 연구에서는 하나의 영상 데이터에서 오디오와 이미지 데이터를 각각 추출하여 학습에 이용하는 멀티 모달 방식 기반의 영상 감정 분류 모델을 제안한다. 실험에는 7가지 감정으로 분류된 Youtube 영상이 사용되었으며, 오디오 데이터는 MFCC(Mel-Frequency Cepstral Coefficient)를 이용하여 영상 당 1개의 오디오 데이터가 추출되었고, 이미지 데이터는 초당 1개의 이미지 데이터를 추출하였다. 각각의 데이터는 Fig. 1과 같이 ViT와 VGGNet(Visual Geometry Group) 모델의 학습에 사용하였으며, 각 데이터를 통해 학습된 ViT, VGGNet 모델을 병합하여 최종적인 영상 데이터 감정 분류 모델을 제작하였다.

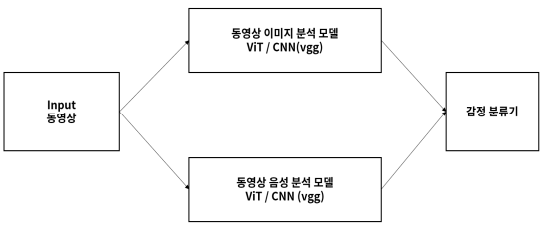

Fig. 1. System Architecture

II. Preliminaries

1. Related works

영상 데이터 감정 분류에 관한 선행 연구의 대부분은 영상에 등장하는 사람의 표정에 의존하여 감정을 분류한다. [2]는 CNN 모델을 이용하여 영상 속 사람의 얼굴 표정을 분류한다. [3]에서 제안한 방법은 멀티 모달 기반의 영상 속 사람의 표정을 기반으로 감정을 분류한다. [4]의 연구에서는 연기 지망생 및 연기 전문가 100명을 대상으로 총 7가지 감정에 대해 1인당 약 100번의 연기를 수행하도록 하여 5초에서 10초 사이의 영상을 촬영하여 얻은 데이터를 사용하였다. 앞선 연구들은 영상에서 사람의 얼굴이 등장할 때 좋은 성능을 기대할 수 있지만, 사람 얼굴이 등장하지 않는 상황에서 영상의 감정을 분류하기 어렵다. 영상에 사람의 얼굴이 등장하지 않고도 영상의 분위기나 장소, 배경 음악으로도 시청하는 사람에게 여러 감정을 전달할 수 있고, 영상의 감정과 영상 내 사람의 표정이 다른 경우도 존재한다. 영상 내 사람의 표정에 의존하지 않고 영상의 감정을 분류할 필요가 있다.

III. The Proposed Scheme

1. Data set

본 논문에서 사용한 데이터 셋은 AIHUB에서 제공된 동영상 콘텐츠 하이라이트 편집 및 설명(요약) 데이터로 동영상 데이터와 하이라이트 구간 및 라벨링 데이터이다. 해당 데이터 셋에서 기타로 분류된 유튜브 동영상 데이터와 '분노', '슬픔', '불안', '상처', '당황', '기쁨', '중립' 7개로 분류된 감정 라벨링 데이터를 사용하였다. 동영상을 라벨링 구간별로 자르고 자른 동영상에 감정을 라벨링하여 데이터 셋을 구축하였다. 다른 감정 클래스와 비교해 기쁨과 중립에 지나치게 많은 데이터가 존재하는 클래스 불균형이 존재했으며, 이가 학습에 악영향을 끼칠 수 있어 Table 1과 같이 총 2941개로 조정하여 데이터 클래스의 비율을 맞춰주었다. 총 9.2GB의 2928개의 동영상을 학습 데이터, 검증데이터, 실험 데이터 7:2:1의 비율로 나누어 학습을 진행하였다.

Table 1. Dataset

감정	데이터 수
분노	265
슬픔	461
불안	400
상처	48
당황	584
기쁨	595
중립	588
총 영상 데이터 수	2941

2. Experimental methods and results

동영상의 감정을 분류하기 위해 동영상의 이미지 데이터와 오디오 데이터를 모두 사용하였다. 동영상에서 이미지 데이터를 추출하여 감정을 분류하는 이미지 분류 모델 학습에 사용하였고, 오디오 데이터를 추출하여 감정을 분류하는 오디오 분류 모델 학습에 사용하였다. 분류 모델로 CNN기반의 모델 중 가장 성능이 좋은 VGG모델과, Transformer기반의 ViT모델을 선택하여 성능을 비교, 분석하였다. 과적합 방지를 위해 검증데이터의 loss값을 기준으로 early stopping을 실행하여 학습을 중단시켰고, 가장 좋은 모델을 선택해 이미지 분류 모델과 오디오 분류 모델을 두 가지 방식의 결합 모델을 이용하여 결합하였다.

2.1 Image classification model

이미지 분류 모델은 동영상의 이미지 데이터를 사용해 감정을 분류하는 모델이다. 이미지 분류 모델로 사전 학습된 VGG모델과, ViT모델의 성능을 비교, 분석하였다. 두 모델 모두 사전 학습된 모델을 사용했고 학습 환경은 동일하게 진행되었다.

이미지 분류 모델의 학습 데이터로 사용하기 위해 구축한 데이터 셋의 동영상에서 1초 단위로 (224, 224)크기의 이미지를 추출하였다. 총 50280장의 이미지를 추출하였고, 학습 데이터 35036장, 검증데이터 10224장, 실험 데이터 5020장으로 나누어 학습에 사용하였다. 배치사이즈는 16으로 설정했고, optimizer는 adamW을 사용했다. Early stopping 기준은 검증데이터의 loss값으로 설정했고, patience는 30으로 설정했다. 각 epoch마다 모든 모델을 저장해 주었다. VGG 모델은 32개의 모델이 저장되었고, ViT 모델은 39개의 모델이 저장되었다.

저장된 모델들로 실험 데이터를 사용하여 동영상 분류 성능을 확인하였고, 방법은 다음과 같다. 먼저 동영상에서 1초 단위로 추출된 이미지들을 모델의 입력값으로 넣는다. 이미지별로 각 7개의 감정으로 예측할 확률값들을 추출한다. 모든 이미지의 확률값들을 다 더해서 가장 높게 예측한 감정을 동영상 분류 결과로 정한다. 이 방법을 사용하여 동영상 분류 결과를 확인해 봤을 때 VGG모델은 19번째 epoch에서 저장된 모델이 정확도 0.46로 가장 성능이 좋았고, ViT모델은 8번째 epoch에서 저장된 모델이 정확도 0.43로 가장 성능이 좋았다.

2.2 Audio classification model

오디오 분류 모델은 동영상의 오디오 데이터를 사용해 동영상의 감정을 분류하는 모델이다. 오디오 분류 모델도 이미지 분류 모델과 동일하게 사전 학습된 VGG 모델과 ViT 모델을 사용하여 비교, 분석하였다.

오디오 분류 모델의 학습데이터로 사용하기 위해 구축한 데이터 셋의 동영상의 오디오 데이터를 16000Hz의 MFCC 특징값으로 추출하여 이미지로 변환하였다. 총 2941장의 이미지를 학습 데이터 2049장, 검증 데이터 584장, 실험 데이터 295장으로 나누어 학습에 사용하였다. 배치사이즈는 60으로 설정했고, optimizer는 AdamW를 사용

했다. Early stopping 기준은 검증데이터의 loss값으로 설정했고, patience는 30으로 설정했다. 각 epoch마다 모든 모델을 저장해 주었다. VGG 모델은 42개의 모델이 저장되었고, ViT 모델은 35개의 모델이 저장되었다.

저장된 모델들의 성능은 실험 데이터를 사용해 확인하였다. VGG모델은 33번째 epoch에서 저장된 모델이 정확도 0.35로 가장 성능이 좋았고, ViT모델은 26번째 epoch에서 저장된 모델이 정확도 0.34로 가장 성능이 좋았다.

2.3 Merge models

이미지 분류 모델과 오디오 분류 모델의 병합 방법은 Fig. 2와 같이 ①데이터를 병합, ②분류기 이후 결과 Max()라는 두 가지 방식으로 실험을 진행하였다.

①데이터를 병합하는 방법은 동영상에서 추출한 이미지 데이터들과 이미지로 변환한 오디오 데이터를 병합하여 학습 데이터로 사용하는 방법이다. 이 경우에도 앞서 학습한 방법과 같이 VGG모델과 ViT모델 두 가지를 학습에 사용하였다. Table 2는 병합 결과로, 각 모델의 감정 분류 정확도와 모델의 정확도를 표로 나타내었고, Fig. 3의 각각 VGG, ViT의 결과를 나타낸 표이다. ViT 기반 모델의 정확도는 45%, VGG 기반 모델의 정확도는 48%로 나타났다.

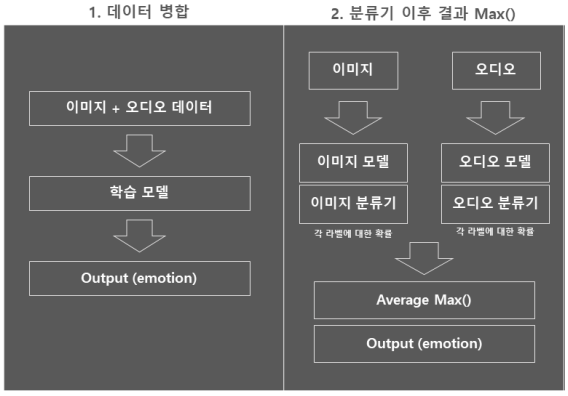

Fig. 2. Merge models

Table 2. Merge model result

	VGG	ViT
감정	Accuracy(%)	Accuracy(%)
분노	39.3	39.3
슬픔	57.5	45.0
불안	56.4	38.5
상처	28.6	28.6
당황	55.9	52.2
기쁨	35.7	48.2
중립	49.2	44.4
	48	45

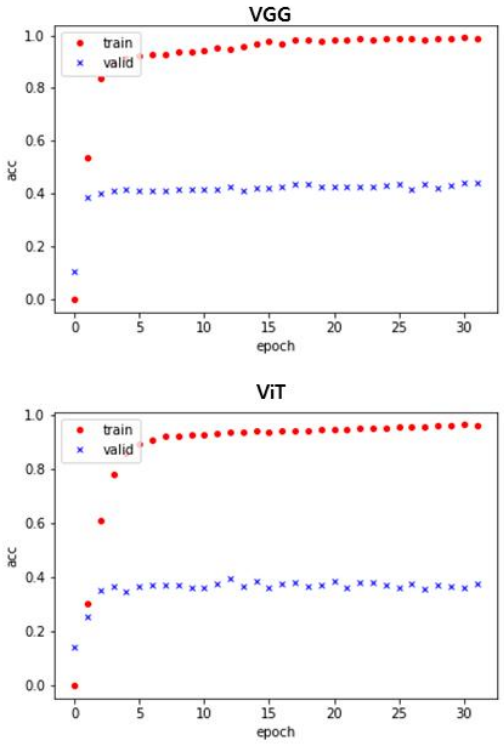

Fig. 3. Merge model result

②분류기 이후 결과 Max() 방법은 다음과 같다. 오디오 분류 모델과 이미지 분류 모델 각각에서 가장 좋은 성능을 보인 모델을 불러와 각각의 데이터에 대한 확률 라벨을 구하고, 각각 7개의 감정 클래스에 대한 확률로 이루어진 (1, 7) 사이즈의 텐서를 평균 내어 새로운 (1, 7) 사이즈의 텐서를 얻어, 가장 큰 값을 가지는 라벨을 아웃풋으로 얻는다. 병합 결과 ViT 기반 모델의 정확도는 30%, VGG 기반 모델의 정확도는 41%로 나타났다.

IV. Conclusions

본 논문에서는 멀티 모달(오디오, 이미지) 기반 ViT 모델을 이용하여 동영상 emotion을 분류하였다. 추후 연구에서는 본 연구에서 학습 성능에 가장 크게 영향을 주었다고 판단되는 데이터의 양을 늘리고 7개의 감정을 긍정, 부정, 중립으로 통합하여 분류 정확도를 높이기 위한 연구를 진행할 예정이다.

ACKNOWLEDGEMENT

이 연구는 과학기술정보통신부의 재원으로 한국지능정보사회진흥원의 지원을 받아 구축된 "동영상 콘텐츠 하이라이트 편집 및 설명(요약) 데이터"를 활용하여 수행된 연구입니다. 본 연구에 활용된 데이터는 AI 허브(aihub.or.kr)에서 다운로드 받으실 수 있습니다.

REFERENCES

[1] Alexey., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale” Computer Vision and Pattern Recognition (2020)

[2] Wisal Hashim Abdulsalam., “Facial Emotion Recognition from Videos Using Deep Convolutional Neural Networks” International Journal of Machine Learning and Computing, Vol. 9, No. 1, February (2019)

[3] 방재훈, 임호준, 이승룡. “실시간 동영상 스트리밍 환경에서 오디오 및 영상기반 감정인식 프레임워크” 한국정보처리학회 춘계학술발표대회 , (2017) : 1108-1111

[4] 문석호, 김성범. "한국어 영상 데이터 감정 분류를 위한 멀티 모달 딥러닝 모델" 대한산업공학회 추계학술대회 논문집 , (2020) : 2944-2955.