감성 분석의 다양한 모델들에 대한 실용성 평가

CUAI 동계 컨퍼런스 4 기 holliday 팀
김민구(기계공학과), 송유지(소프트웨어학부),신재현(컴퓨터공학), 유승태(컴퓨터공학), 이민정(전기전자공학부)

[요약]
상황에 따라 행동이나 말을 결정하고 그 속에서 묻어나는 감정을 파악하는 것은 인간 고유의 영역이었지만, 인공지능 기술이 발전됨에 따라 기계가 사람의 감정을 파악하는 기술 또한 더 이상 불가능한 일이 아니게 되었다. 이러한 기술은 마케팅에 대한 활용도가 높고, 소셜 미디어의 사회적 부작용이 심각해지는 상황에서 자살암시 테러계획등과 같은 사람의 감정을 파악하는 등에 활용 가능성이 높아 필요성이 대두되고있다. 그에 따라 감정 분석은 다양한 데이터를 활용한 많은 분야에서 현재 실제 생활에 적용될 수 있을만큼 많은 발전을 이루어낼 수 있었다. 우리 팀은 위와 같은 감정 분석에 관련한 다양한 접근 방식들을 직접 현실 세계를 보다 잘 반영한 데이터를 통해 이들 기술이 얼마나 발전되었는지 파악하고, 그리고 어떤 부분에 있어 아직 발전 가능성이 존재하는지 분석한다..

1. 서 론

인간이 감정을 파악하는 것에 있어 가장 많이 사용되는 데이터는 표정, 어조, 발화 내용 등이다. 따라서 우리는 이미지, 음성, 텍스트를 활용한 감정 인식 방법에 중점을 두고 조사 및 테스트를 진행하였다.

이미지를 활용한 분석에서는 기본적인 Deep Convolutional Neural Network 를 활용한 연구부터 시작하여, 이들만으로 해결할 수 없는 표정의 불명확함을 해결할 수 있는 새로운 방법론을 조사하였다.

텍스트 감성 분석의 경우 BERT 를 사용하여 진행한 연구에 대해 주로 조사하였다. BERT 를 사용하여 관련 대량 코퍼스를 Encoder 가 임베딩 하고, 이를 전이 학습 및 파인튜닝한 후 분류를 원하는 데이터에 대해 머신러닝 모델을 이용하여 분석을 진행한다. BERT 는 특정 분야에 국한된 기술이 아니라 모든 자연어 처리 분야에서 좋은 성능을 내는 범용 Language Model 이고 지금까지 자연어 처리에 활용하였던 앙상블 모델보다 더 좋은 성능을 내고 있어서 많은 관심을 받고 있는 언어 모델이다.

음성 데이터를 통한 감성 분석은 머신러닝과 딥러닝의 기술 발전과 더불어 빠르게 발전하고 있으며, 현재 많은 음성 분류 연구는 Mel Frequency Cepstral Coefficient 와 Mel Spectrogram 을 활용한 RNN/CNN 기반 모델들을 사용하고 있다.

다양한 모델들의 실용성을 확인하기 위해 기존의 벤치마크를 위한 데이터를 활용하는 것은 실제 유효성을 확인하기 어렵다고 판단되어 AI-HUB 의 '멀티모달 영상 데이터셋'을 활용하였다. 해당 데이터셋은 110 시간 분량의 총 6 천개 영상 클립으로 구성되어 있고 클립당 1-3 분 내외의 대화를 포함하는 영상 파일과 영상을 설명하는 메타파일로 구성되어 있다. 발화 정보에는 학습 과정에서 정답 라벨로 활용할 수 있는 감정과 발화별 대화 의도 및 대화전략 정보가 부착되어 있다.

2. 본 론

본론에서는 각각의 방법에 대한 데이터 처리 과정 및 모델링 방법에 대해 기술 하려한다.

1. 텍스트를 활용한 감정 분석

텍스트를 활용한 감정 분석을 진행하기 위해 json 데이터 전처리, 감정 라벨 인코딩, 감정 분류 모델 생성을 진행하였다.

1.1 json 데이터 전처리

AI hub 의 각 영상별 .json 파일에는 사람의 감정 정보 이외에도 사람 객체 정보, 상황 정보 등 해당 영상과 관련된 다양한 정보들이 포함되어 있다. 영상별 json 에서 각 프레임 별로 감정 정보가 포함된 부분의 key 만 남기는 작업을 진행하였다.

감정 정보에는 텍스트에 대한 감정 정보 이외에도 이미지와 음성에 대한 감정 정보 또한 포함되어 있다. 텍스트를 활용한 감정 분석을 진행하기 위해 각 영상 json 파일의 프레임 별 감정 정보 중 텍스트 데이터에 대한 감정 정보와 해당 프레임에서의 한국어 script 를 뽑아 csv 파일로 저장하였다.

각 영상 csv 파일에서 script 의 중복이 존재하여 중복을 제거하는 작업을 진행하였다. 그리고 각 영상별 csv 파일을 하나의 feature 파일로 통합하였다.

1.2 감정 라벨 인코딩

영상 클립의 프레임에 부착된 데이터 예시를 보면 분노(anger), 경멸(contempt), 혐오(dislike), 공포(fear), 행복 (happy), 슬픔(sad), 놀람(surprise), 중립(neutral) 의 감정 라벨 등이 부착되어 있다. 중립(neutral) 감정의

경우 분류하기 애매하기 때문에 사용하지 않고 나머지 감정의 script 를 사용하였다.
angry, contempt, dislike, fear, happy, sad, surprise 순서대로 라벨 인코딩을 진행하였다. Sklearn 모듈의 LabelEncoder 메서드를 활용하였다.

1.3 감정 분류 모델 생성

감정 분류 모델로는 ktrain 오픈소스 라이브러리를 활용하였다.

ktrain 의 Transformer[9] class 및 albert-large[10] model 의 pretrained weight 을 이용하였다.

감정 분류 모델 생성 과정은 다음과 같다.

1. ktrain.text.Transformer 객체 생성 (maxlen = 64)
2. train 과 test 데이터에 대한 preprocessing
3. 감성 분석을 위한 분류 모델을 생성
4. keras model 을 tune 하고 train 하는데 이용되는 learner instance 를 생성 (batch size = 64)
5. 학습을 통해 최적의 learning rate 를 찾음
6. 찾은 learning rate 값을 이용하여 learner 를 학습

학습시킨 모델을 활용하여 입력한 데이터에 대한 감성 분석을 진행하려면 predictor 를 생성하면 된다.

1.4 성능 및 추후 진행 방향

Script 텍스트와 감정 라벨을 이용하여 학습을 진행한 결과 36% 정도로 다소 아쉬운 정확도를 기록하였다. loss 값이 높은 데이터들을 뽑아서 감정 라벨과 모델 예측 감정을 비교해 본 결과 모델 예측 감정이 오히려 해당 문장에 적절한 감정인 경우도 있었다. 라벨링이 잘못되어 감정을 잘못 예측하는 경우도 있으므로 성능 향상을 위해서는 학습 데이터 감정 라벨링 등의 추가적인 전처리가 필요해 보인다. 또한 사용한 데이터의 경우 감정별 데이터 개수가 불균형하다. 성능 향상을 위해서는 각 감정마다 충분한 학습 데이터를 확보할 필요가 있다.

	precision	recall	f1-score	support
fear	0.37	0.13	0.19	175
angry	0.00	0.00	0.00	74
contempt	0.31	0.03	0.05	393
sad	0.42	0.56	0.48	685
surprise	0.29	0.65	0.40	474
dislike	0.37	0.29	0.32	435
happy	0.58	0.23	0.33	310
accuracy			0.36	2546
macro avg	0.33	0.27	0.25	2546
weighted avg	0.37	0.36	0.32	2546

Fig 1. ktrain ALBERT 모델 결과

```
array([[ 22,   0,   4,  45,  81,  16,   7],
       [  5,   0,   3,  23,  37,   6,   0],
       [ 14,   0,  11, 101, 226,  35,   6],
       [  5,   0,   5, 382, 172,  86,  35],
       [  8,   0,  10,  95, 307,  51,   3],
       [  3,   0,   1, 121, 184, 125,   1],
       [  3,   0,   1, 152,  62,  21,  71]])
```

Fig 2. ktrain ALBERT 모델 confusion matrix: 왼쪽 대각선에 해당하는 숫자가 높을 수록 정확도가 높은 모델.

2. 이미지를 활용한 감정 분석

이미지를 활용한 감정분석으로는 다양한 논문들과 함께 다양한 오픈 소스 코드가 존재하는 Kaggle 과 Github 를 활용하였다. 또한 학습을 위하여 FER-2013 데이터셋을 활용하였다.

2.1 데이터 전처리

학습을 위한 데이터는 이미 전처리가 완료된 FER-2013 데이터셋을 활용하였기에 전처리가 필요하지 않았으나, 테스트를 위한 데이터의 경우 학습용 데이터와 유사한 형식을 맞춰야했기에 전처리가 활용되었다.
기본적으로 얼굴에 중심이 맞춰진 학습용 데이터와 유사한 형식을 위해 DLib 에서 제공하는 Face Detector 을 활용하여 얼굴의 위치를 추출하였다. 이후 해당 좌표를 top-left, bottom-right 에서 x, y, width, height 형식으로 변환하여 얼굴의 위치만을 추출하였다. 이때 dlib 의 특성상 음수 좌표, 혹은 범위를 벗어난 좌표를 반환하는 경우도 존재하기 때문에 최대 최솟값을 지정하여 올바른 범위내에 포함되도록 하였다. 이후 RGB 색상은 표정을 인식하는 것에있어 큰 영향을 주지 않기 때문에 gray scale 로 변환 후 48x48 사이즈의 이미지로 크기를 맞춘다.

2.2 테스트

통상적 접근 방식 이미지를 활용한 감정분석이라는 문제에 접근하는 방식 중 가장 일반적으로 확인 가능한 방식들로는 VGG, ResNet 등 과 같이 이미지에 관련된 문제에 좋은 성능을 보이는 모델들을 활용하는 방식이다. 실제로 단순히 ResNet, VGG, InceptionNet 등 과 같은 네트워크를 가용하는 것만으로도 70%이상의 높은 정확도를 확인할 수 있었다. 또한 그보다 간단한 구조를 가진 6 개의 Convolutional Layers 로 이루어진 Deep CNN 을 모델링하여 FER-2013 데이터셋을 학습시켰을때에 70%이상의 높은 정확도를 보였다.
이렇게 학습된 모델을 다른 얼굴 이미지나 AI-hub 의 멀티모달 데이터셋에 적용하였을때 또한 좋은 결과를 보이는 것을 확인하였다. 이러한 방식의 문제점 또한 존재했는데 입의 모양이 감정을 판단하는 것에 있어 큰 영향을 준다는 것과 모호한 감정을 표현하는 얼굴의 경우에 일부 감정에 편향된 결과를 보인다는 것이다.

이러한 문제점들은 신뢰성있는 인공지능을 개발하는 것에 있어 문제가 될 수 있다.

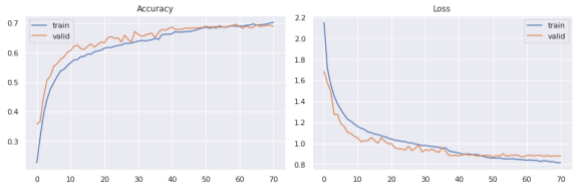

Fig 3. DCNN 모델 결과: (왼쪽) epoch 에 따른 정확도. (오른쪽) epoch 에 따른 손실값.

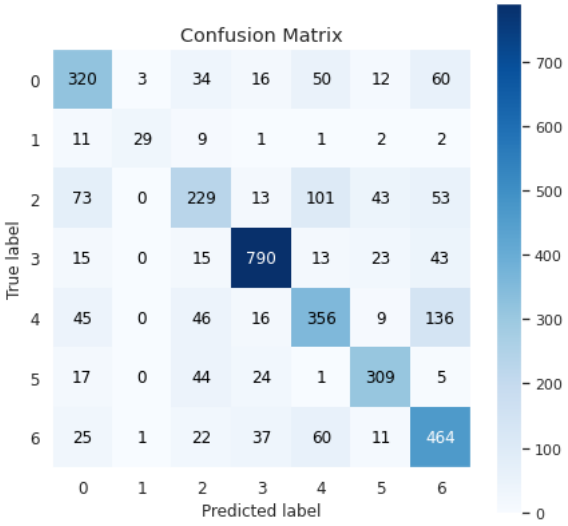

Fig 4. DCNN 모델의 Confusion Matrix

Relative Uncertainty Learning 해당 논문은 NeurIPS 2021 에 게재된 논문으로, 앞에서 언급된 감정의 불확실성을 해결하기 위한 방법을 제시하였다. 저자는 "상대적 불확실성"이라는 개념을 통해 불확실성 문제를 해결하는 방식을 제안한다. 해당 방식은 MixUp 과 유사한 방식을 활용하여 불확실성이 큰 이미지를 중점적으로 학습할 수 있도록 한다.

2.3 추후 방향성

RUL 과 같은 논문 뿐 아니라 복합적인 감정을 잘 구분할 수 있는 방법을 찾기 위해 많은 노력이 이루어지고 있다. 하지만 많은 자료들에서 언급되는 것 과 같이 감정이 명확하게 표현되지 않은 표정들은 사람의 눈으로도 판단하기 매우 어렵기 때문에 보다 다양한 방식의 접근이 필요할 것이라고 판단된다.

RUL 에서 소개된 것과 같이 불확실성을 집중적으로 해결하는 방안 또한 중요할 것이라 예상되지만, 라벨링 자체에서도 문제가 존재한다고 보여진다. 하나의 표정이 사람의 눈으로도 판단하기 어려운 이유로는 해당 표정에 복합적인 감정이 포함되어있기 때문이다. 따라서 기존의 hard label 을 활용한 방식은 표정의 복잡성을 표현하는 것에있어 충분치 않다. 이를 해결가능한 방안으로는 soft label 을 사용하는 방법 등이 존재한다.

3 음성을 활용한 감정 분석

3.1 데이터 전처리

AI hub 의 영상 데이터(.mp4)를 먼저 음성 파일로 변환한 후에 .json 파일을 통해 영상 내 각 대화의 시작 프레임과 끝 프레임을 가지고 음성 파일을 자르고 emotion label 을 가져온다. 변환된 음성 파일(.wav)은 모델에 들어가기 전에 전처리를 거쳐야 한다.

음성 파일들은 mono(1 개 채널) 또는 stereo(2 개 채널)이며, 대부분 stereo 이다. 차원을 같게 하기 위해서 mono 파일들을 stereo 파일로 변환해줘야한다. 또한, sampling rate 도 음성 파일마다 다르기 때문에 (대부분 44100Hz) 모델에 입력되는 차원이 같도록 하나의 통일된 sampling rate 으로 standardize 해야한다. 영상 내 각 대화별로 음성 데이터를 잘랐기 때문에 각 음성 데이터 파일의 길이는 서로 다르다. 음성 샘플들의 길이가 서로 같게 만들기 위해서 zero padding 을 실시한다. 데이터 augmentation 기법으로는 총 4 가지를 선택했다 (pitch change, shifting, speed change, white noise 추가).

이렇게 전처리를 한 오디오 데이터는 Mel Frequency Cepstral Coefficient (MFCC) 또는 Mel Spectrogram 으로 변환을 해서 모델에 넣는다. MFCC 는 음성 데이터를 feature vector 로 바꿔주는 알고리즘이며, Mel Spectrogram 은 음성 신호를 짧게 잘라서 푸리에 변환(Short Time Fourier Transform)과 L2 norm 을 적용한 후 mel-filter 를 적용하여 얻는다.

음성 모델은 Mel Spectrogram 을 활용하여 bi directional CRNN 모델을 기반으로 학습을 시도했으나 오디오에 대한 지식부족과 시간 부족으로 인해 원활한 학습을 하지 못했다. 그러므로 지금까지 음성 데이터로 모델을 구축하고 학습하는 것은 앞으로의 과제로 남을 것이다.

3. 결론

다양한 도메인의 데이터를 활용한 테스트와 최근 동향에 관한 조사를 통하여 해당 방식들을 통하여 유의미한 결과를 얻어낼 수 있고 앞으로도 추가적으로 발전되어야할 부분이 존재한다는 것을 확인할 수 있었다.

텍스트의 경우 대량의 코퍼스(말뭉치)를 사전 학습한 후 해당 사전 학습 weight(e.g. BERT 언어 모델 출력)에 추가적인 신경망을 쌓아 원하는 구체적인 감성 분석 task 에 적용하는 기술이 연구되고 있다.

이미지는 불확실성을 잘 이해하는 모델 구조를 통해 미묘한 감정들을 잘 파악 가능한 기술들이 연구되고 있다.

공통적으로 보여지는 문제점은 복합적인 감정을 표현할 수 없는 라벨링으로 보여진다. 멀티모달 데이터셋에서도 표정, 음성, 텍스트가 모두 다른 감정으로 라벨링되어있는

경우도 심심치 않게 보여진다. 따라서 복잡한 감정을
학습할 수 있도록하는 라벨링의 필요성이 보여진다.

참고 문헌

[1] "음성,이미지, 텍스트를 동시 인식하는 AI 플랫폼 설명 ". Available: http://www.aitimes.kr/news/articleView.html?idxno=24067

[2] "Github Link for Facial Expression Recognition". https://github.com/WuJie1010/Facial-Expression-Recognition.Pytorch

[3] Zhou Yue, Feng Yanyan, and Zeng Shangyou "Facial Expression Recognition Based on Convolutional Neural Network", IEEE International Conference on Software Engineering and Service Sciences, Oct. 2019

[4] Kai Wang, Xiaojiang Peng, and Yu Qiao, "Suppressing Uncertainties for Large-Scale Facial Express Recognition" Computer Vision and Pattern Recognition (CVPR), Nov. 2020

[5] "Image Sentiment Analysis" . Available: https://github.com/zyh-uaiaaaa/relative-uncertainty-learning

[6] AI-HUB 데이터 출처 "멀티모달 영상" https://aihub.or.kr/aidata/137

[7] Yuhang Zhang, Chengrui Wang, and Weihong Deng "Relative Uncertainty Learning for Facial Expression Recognition" 35th Conference on Neural Information Processing Systems (NeurIPS 2021).

[8] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[9] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems. 2017.

[10] Lan, Zhenzhong, et al. "Albert: A lite bert for self-supervised learning of language representations." arXiv preprint arXiv:1909.11942 (2019).

[11] "이미지(표정) 인식에 활용한 kaggle 소스" https://www.kaggle.com/aayushmishra1512/emotion-detector