CNN을 활용한 음성 감정 분석 및 화자 식별

2021741062 김현준

HyunJun Kim

딥러닝을 통해 산업계의 문제를 어떻게 해결할 수 있을지 생각해본 결과, 특정 직업의 처우 개선에 딥러닝을 사용할 수 있을지에 대해 고민해보게 되었다. 다양한 직업들 중 사람들과 많이 마주하는 직업이 스트레스를 많이 받는 다는 것을 보았고, 이에 사람들과 많이 마주한다고 판단한 상담사 분들의 관점에서 딥러닝을 어떻게 사용할 수 있을까에 대해 고민하였다. 고객들과 전화를 할 때 현재 상대방의 감정을 실시간으로 파악할 수 있다면, 그리고 상대방이 누군지 알 수 있게 된다면 상담에 효율과 직업 만족도를 상승시킬 수 있지 않을까 라는 생각으로 프로젝트를 시작하였다.

1. 서 론

4학년으로 올라오며 다양한 업체와 컨택하고 연락하는 과정 속에서 전화를 할 일이 굉장히 많았다. 각 업체에 상담사 분이 한 명만 계신다면 전화를 할 때 바로 본론으로 넘어갈 수 있겠지만, 실제로는 그렇지 않았기에 매번 내가 누군지 설명하고, 지난 상담에서는 어디까지 말하였고, 오늘은 무엇을 위해 전화를 드렸는지 설명하는 과정이 필요했다.
또한 전화를 지난번 상담사 분과 다른 분이 받으시게 되면, 내가 지난 상담에서 어디까지 얘기를 했다고 전해드려도 잘 이해를 못하시고 답답해하시는 것을 느낄 수 있었다. 이에 상담사 분들이 다양한 문의를 받는 과정에서 딥러닝을 활용하면 업무의 강도나 효율성을 높일 수 있을 것이라고 생각하였다.

2. 본 론

내가 만약 상담사의 일을 수행하게 되었을 때, 어떤 부분이 가장 중요한가? 라고 생각한다면 그것은 바로 **감정분석**이다.
여기서 말하는 감정분석은 고객의 감정과 본인의 감정을 모두 포함하는 부분이다. 먼저 고객의 감정 분석이다. 고객의 감정 변화를 모니터링 할 수 있게 된다면 상담사가 적절한 대응을 할 수 있게 된다. 예시로 물건의 잦은 고장으로 화가 나서 전화한 사람에게는 공감의 말이나 안타까움을 표현할 수 있을 것이다. 상담사 본인의 감정 분석은 업무의 강도나, 스트레스 분석을 위한 척도로 상담 도중 본인의 감정 분석 결과가 좋지 않게 뜬다면 잠깐 휴식을 취하는 등의 조치를 취하게 될 수 있을 것이다.

그리고 만약 반복적으로 전화를 하는 사람이 있다고 할 때, 목소리 만으로 그 사람이 누군지 파악할 수 있게 된다면 그 또한 굉장히 큰 이점을 가질 수 있을 것이라고 판단했다. 특정 고객이 다시 전화를 걸 때 이전 상담 기록을 빠르게 조회할 수 있게 되고, 이를 통해 더 빠르고 정확한 응대가 가능해진다. 또한 가끔 전화를 하다 보면 본인 확인을 위해 주소나, 이름을 물어보는데 사실 이 부분은 충분히 다른 사람도 말할 수 있는 정보이기 때문에 목소리를 사용하여 고객을 조회할 수 있게 된다면 보안에도 장점이 있을 것이라고 생각하였다.
그러한 이유로, 이번 프로젝트에서는 음성 데이터를 활용하여 감정을 분석하고, 추가적으로 화자를 식별하는 과정을 수행해보고자 한다.

2.1 데이터 셋 선정

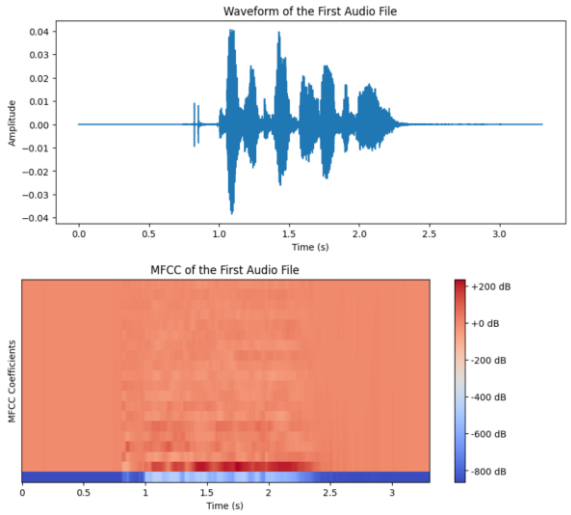
학습에 필요한 음성 데이터를 선정하는 과정도 쉽지는 않았다. 다양한 음성 데이터 세트를 찾아본 결과 RAVDESS, CREMA-D, TESS, EMO – DB, IEMOCAP 등 다양한 데이터를 찾을 수 있었다. 이중에서 화자 식별을 위해 적합한 세트로는 RAVDESS와 CREAMA – D가 있었다. 두개의 데이터 세트 중 결국 RAVDESS를 사용하게 되었는데, 그 이유로는 첫번째, 12명의 사람들이 60번씩 녹음하였기에 학습에 적합하다고 판단했으며, 두번째로는 성우가 모두 미국인이었기 때문이다. CREAMA-D의 데이터 세트는 다양한 인종 및 성별의 배우들이 감정을 표현하기 때문에 인종에 차이에 따른 목소리의 다름이 존재할 수 있을 것이라고 생각하였다. 비슷한 지역에 사는 사람들의 목소리도 구분이 될지가 궁금했기 때문에 RAVDESS 데이터를 사용하였다.

2.1.1 RAVDESS 전 처리

KAGGLE에서 확인할 결과, 파일명에 감정, 발화한 내용, 성별, 성우의 번호까지 다 기록이 되어 있음을 확인하였다. 감정과 성우를 모두 예측하기 위해 감정이 적혀져 있는 3번째 인덱스와, 마지막 인덱스를 예측하는 Y로 포함시켰다. 배우의 경우에는 EVEN 값이 Female이라 적혀 있었기 때문에 이를 활용하여 male, female로 분류를 하였다. 처음에는 성별에 관계없이 감

정을 분류할 수 있도록 방향을 잡았으나, 남자와 여자의 성별에 따른 음성과 감정의 차이는 무조건 있을 것이라고 판단하였기에 다음과 같이 성별을 나누어 실험을 진행하게 되었다.
데이터를 전 처리할 때 학습 CNN 모델에 맞게 그 형태를 바꾸어 주었어야 했는데, Conv1D, Conv2D에 모두 테스트를 해본결과 Conv1D보다 Conv2D에서 더 높은 성능을 보였다. 이에 따라 Conv2D 모델에 맞게 전처리를 할 수 있도록 하였다.

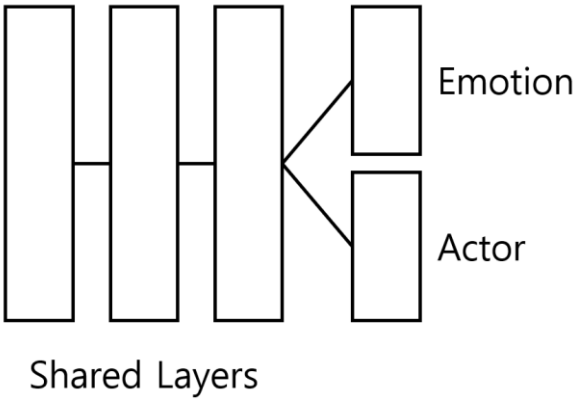
이번 음성 분석을 하기 위해서 입력 받은 wav파일을 MFCC 특징을 추출하여 사용하였다. 이를 위해 librosa 라이브러리를 사용하였다. 오디오 파일을 load하고 librosa.feature.mfcc 함수를 통해 MFCC를 추출하고, 차원을 추가하여 Conv2D 형태에 맞게 바꾸었다. 이때 MFCC가 무엇인지 궁금할 수도 있는데, MFCC란 오디오 신호에서 추출할 수 있는 Feature로 소리의 고유한 특징을 나타내는 수치로, 음성인식이나 화자 분석에 주로 사용된다.
MFCC는 (Mel-Frequency Cepstral Coefficient) 음성 데이터를 짧은 프레임으로 작게 분할하고, 각 프레임에 고속 푸리에 변환(FFT)을 적용하여 주파수 도메인으로 변환시킨다. 이후 주파수 축을 Mel Scale Filter (인간의 청각 특성을 반영한 주파수 척도) Bank를 적용하여 Mel Spectrum을 구한 후, Cepstral 분석 (정보를 추출할 때 사용)을 통해 구할 수 있다. Mel Scale 변환은 주파수 축을 Mel Scale로 변환하여 사람의 청각 특성을 반영하는데, 이 과정에서 불필요한 고주파 성분이 감소하여 잡음에 강하다는 장점이 있다.



[Fig. 1] 입력 데이터의 WaveForm과 MFCC 파형 분석

2.1.2 Multi – task Learning

총 Output이 감정 예측과 화자 인식으로 두가지지 때문에 원래는 학습을 두 번 돌리는 것이 올바르지만, 감정 예측과 화자인식 모두 MFCC 특징을 사용하기 때문에 같은 모델을 활용하여 학습을 진행하면 시간을 절약하고 효율적인 데이터 사용이 가능해진다.



[Fig. 2] Multi – task Learning의 구조 설명

3. 실험 결과

Without White-Noise

```
Test Emotion Output Accuracy: 0.4479166567325592
Test Actor Output Accuracy: 0.6909722089767456
9/9 [==============================] - 0s 4ms/step
Actual Emotion: male_disgust, Predicted Emotion: male_disgust
Actual Actor: 3, Predicted Actor: 3
================================================================
Actual Emotion: male_calm, Predicted Emotion: male_calm
Actual Actor: 11, Predicted Actor: 11
================================================================
Actual Emotion: female_calm, Predicted Emotion: female_fearful
Actual Actor: 10, Predicted Actor: 10
================================================================
Actual Emotion: female_calm, Predicted Emotion: female_calm
Actual Actor: 2, Predicted Actor: 2
================================================================
Actual Emotion: male_angry, Predicted Emotion: male_disgust
Actual Actor: 11, Predicted Actor: 11
================================================================
Actual Emotion: male_calm, Predicted Emotion: male_neutral
Actual Actor: 17, Predicted Actor: 17
================================================================
Actual Emotion: female_calm, Predicted Emotion: female_calm
Actual Actor: 24, Predicted Actor: 6
================================================================
Actual Emotion: female_fearful, Predicted Emotion: female_fearful
Actual Actor: 24, Predicted Actor: 14
================================================================
Actual Emotion: female_surprised, Predicted Emotion: female_fearful
Actual Actor: 20, Predicted Actor: 14
================================================================
Actual Emotion: male_calm, Predicted Emotion: male_disgust
Actual Actor: 5, Predicted Actor: 5
================================================================
```

[Fig. 3] White Noise 없이 학습 완료된 모델의 분석 결과

With White-Noise

```
Test Emotion Output Accuracy: 0.4479166567325592
Test Actor Output Accuracy: 0.3680555522441864
9/9 [==============================] - 2s 237ms/step
Actual Emotion: male_disgust, Predicted Emotion: male_disgust
Actual Actor: 3, Predicted Actor: 7
======================================================================
Actual Emotion: male_calm, Predicted Emotion: male_calm
Actual Actor: 11, Predicted Actor: 15
======================================================================
Actual Emotion: female_calm, Predicted Emotion: female_sad
Actual Actor: 10, Predicted Actor: 5
======================================================================
Actual Emotion: female_calm, Predicted Emotion: female_sad
Actual Actor: 2, Predicted Actor: 2
======================================================================
Actual Emotion: male_angry, Predicted Emotion: male_angry
Actual Actor: 11, Predicted Actor: 17
======================================================================
Actual Emotion: male_calm, Predicted Emotion: male_fearful
Actual Actor: 17, Predicted Actor: 17
======================================================================
Actual Emotion: female_calm, Predicted Emotion: female_calm
Actual Actor: 24, Predicted Actor: 24
======================================================================
Actual Emotion: female_fearful, Predicted Emotion: female_happy
Actual Actor: 24, Predicted Actor: 4
======================================================================
Actual Emotion: female_surprised, Predicted Emotion: female_fearful
Actual Actor: 20, Predicted Actor: 10
======================================================================
Actual Emotion: male_calm, Predicted Emotion: male_calm
Actual Actor: 5, Predicted Actor: 19
======================================================================
```

[Fig. 4] White Noise를 적용한 후 학습 완료된 모델의 분석 결과

MFCC가 외부 노이즈에 강하다는 것을 검증해보기 위해 생성할 수 있는 노이즈들 중 가장 널리 퍼진 White – Noise를 오디오 파일에 추가하여 학습을 진행해보았다. 수행 결과 감정 분석은 비슷한 결과로 나왔지만, 화자 분석은 정확도가 많이 떨어졌음을 확인할 수 있었다.

4. 결 론

MFCC와 CNN을 활용하여 감정분석과 화자인식을 하는 모델을 만들고, 직접 사용해본 결과 분류가 올바르게 진행되고 있음을 알 수 있었다. 감정의 경우 50퍼센트 정도로 나오고 있는데, 소분류가 8가지나 되어 있었기 때문에 (neutral, calm, happy, sad, fearful, disgust, surprised) 제공된 1440개의 오디오 데이터로는 충분한 학습이 진행되기 어려웠다고 생각한다. 8가지 데이터가 아닌 긍정(GOOD), 부정(BAD) 식의 이진 분류를 진행했다면 더 높은 결과를 얻을 수 있겠으나, 사람의 감정을 GOOD, BAD 두개로 나눌 수는 없었기에 기존에 분류된 틀을 사용하여 학습을 진행했다. 화자 분석의 경우 완전 일치하게 분석한 경우도 있었으나, 그렇지 않은 경우는 정확한 화자를 인식하지는 못했지만 화자의 번호로 보았을 때 남녀의 구분은 확실하게 하고 있음을 알 수 있었다.

감정 분석과 화자 분석을 같은 전처리를 거쳐 사용했지만, 사실 둘은 같은 전처리를 사용하기에는 그 한계가 있을 것이라고 생각한다.

감정 분석의 경우 음성에서 화자의 감정을 예측하기에 음성 신호의 패턴과 주파수의 변화를 사용한다. 주파수 스펙트럼의 경우 분노한 음성은 더 높은 주파수를 띄게 되며, 강하고 높은 억양은 보통 기쁨이나 놀람을 표시하게 된다. 그랬기에 MFCC로 전처리한 감정 분석은 노이즈에 큰 영향을 받지 않는 모습을 확인할 수 있었다.

화자 분석의 경우에는 음성에서 화자를 식별하는 것을 목표로 하며, 주파수의 변화 보다는 목소리의 음색, 발음 패턴 등을 사용해서 학습을 해야 한다. 그랬기에 주파수 데이터를 기반으로 한 특징인 MFCC만으로는 노이즈에 강인한 모델을 만들 수 없었다고 생각한다. 다른 특징들을 가지고 전처리를 했다면 더 강인한 모델을 만들었을 수 있을 것이라고 생각한다.

또한 MFCC는 목소리의 주파수를 바탕으로 감정을 파악하는 것이기에 말의 뜻을 해석하지는 못한다. 예를 들자면 남편의 (자기야 플스 사도 돼?) 라는 질문에 와이프의 (그래 사 ㅎㅎ) 라는 발화만을 본다면, MFCC 분석을 통한 와이프의 발화는 CALM으로 인식을 할 가능성이 크다. 하지만 뜻을 파악한다면 CALM이 아니라 ANGRY일 확률이 크다는 것을 우리는 알 수 있다. 그러하였기에 구글의 음성인식 라이브러리 Speech Recognition과 Bert 모델을 활용하여 학습에 사용된 음성을 텍스트로 변환하고, 만들어진 텍스트도 학습에 사용하여 말의 주파수와 뜻까지 판단에 사용하는 멀티 모달(Multimodal) 모델을 학습시켜 사용하고 싶었으나, 역량의 한계로 그렇게까지 발전시키지 못한 것이 아쉽다.

References

[1] MFCC 이해하기, [Online], Available : https://brightwon.tistory.com/11

[2] MEL Spectrogram, MFCCs, Chroma Frequencies, [Online], Available: https://sswwd.tistory.com/4

[3] Multi – task Learning, [Online], Available: https://mapadubak.tistory.com/40