CNN-GRU모델을 이용한 음성데이터 감정 분석 연구

서경빈, 이동욱, 김남호, 최광미*

*호남대학교 컴퓨터공학과

{tjrudqls1003, irum1020}@naver.com, {nhkim,*cgmi66}@honam.ac.kr

Research on emotion analysis of voice data using CNN-GRU model

Gyeongbin Seo, Dongwook Lee, Namho Kim, Gwangmi Choi*

*Dept. of Computer Engineering, Honam Univ.

요 약

사람들 간의 의사소통에서 감정을 파악하는 것은 중요하다. 이 논문은 음성 데이터를 활용하여 감정을 정확히 인식하고, 그에 맞게 대응하기 위해 CNN을 통해 공간적인 특징을 추출하고 GRU를 통해 감정 변화를 학습하는 CNN-GRU 모델을 제안한다. 이 모델은 짧은 시간 간격 내의 음성 데이터에 적합하며, 메모리 사용량이 적고 효율적이다 더불어, CNN-GRU 모델을 결합하여 두 계층의 분석을 통해 보다 정확한 감정 표현을 목표로 구현하고자 한다.

Ⅰ. 서 론

감정은 사람의 마음을 표출할 수 있는 가장 중요한 요소라 할 수 있다. 이런 인간의 감정은 오랜 시간을 거쳐 심리학자와 신경 과학자들에 의해 인간의 감정과 표현 방식을 이해하고 분류하는 연구를 해 왔다[1]. 사람들은 대화 중 상대방의 감정을 인식하고 이에 맞게 대응한다. 따라서 음성 데이터에서 감정을 정확하게 인식하는 것은 사회적 상황을 올바르게 인식하는 데 도움이 된다. 음성 감정인식이란 음성신호를 분석하여 화자의 감정을 자동으로 인식하는 기술이다. 음성을 이용한 감정인식은 데이터의 취득이 간편하다는 장점이 있지만 다른 방법에 비해 성능이 낮다는 단점이 있다[2]. 본 논문은 DCNN(Deep Convolutional Neural Network)알고리즘의 CNN모델과 GRU모델을 결합하여 두 계층의 분석으로 보다 정확한 감정을 분석하여 활용하고자 한다.

Ⅱ. 관련연구

<표 1>은 IEMOCAP(Interactive Emotional Dyadic Motion Capture)을 활용하여 다양한 감정을 측정하는 맥락에서 딥 러닝, 즉 DCNN(Deep Convolutional Neural Network) 알고리즘과 기존 알고리즘을 상세히 비교 분석한 것이다. 이 연구에서는 Emo-DB와 SAVEE를 데이터셋으로 사용하여 행복, 분노, 슬픔 등 다양한 감정을 인식하였다[3]. 본 논문에서는 기존 기술에 비해 감정 인식에서 우수한 성능을 보여주는 CNN모델과 GRU 모델을 결합하여 두 계층의 분석으로 보다 정확한 감정을 분석하여 활용하고자 한다.

표 1. 다양한 분류기의 비교 분석

Algorithms	Anger	Happy	Sad
k-nearest neighbor	93%	55%	77%
Linear discriminant analysis	68%	49%	72%
Support vector machine	74%	70%	93%
Regularized discriminant analysis	83%	73%	97%
Deep Convolutional neural network	99%	99%	96%

CNN(Convolutional Neural Network)과 LSTM(Long Short-Term Memory)을 조합한 기존의 연구를 살펴보고자 한다. 음성 스펙트로그램에서 특징을 추출하기 위해 CNN을 사용하고, LSTM 신경망을 활용하여 시간 정보를 고려하여 감정을 예측했다. 이 연구에서는 6가지 감정 카테고리를 대상으로 실험을 진행했으며, 최종적으로 가중 정확도는 61%에 이르렀으며, 비가중 정확도는 56%로 나타났다[4].

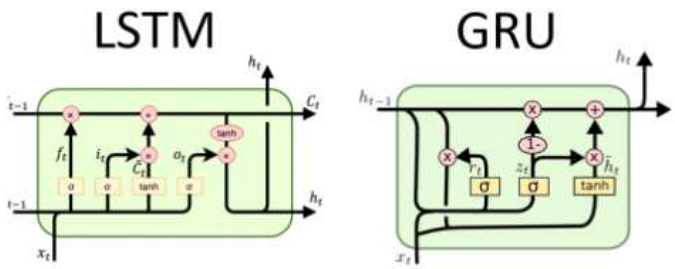

그림 1. LSTM/GRU Architecture <출처: dprogrammer.org>

<그림 1> LSTM 셀은 셀 상태를 유지하면서, 입력게이트, 망각게이트, 출력게이트를 이용하여 출력값을 조절한다. 입력게이트는 입력값을 얼마나 받아들일지를 결정하고 망각게이트는 이전의 셀 상태를 얼마나 잊어버릴지를 결정하며 출력게이트는 얼마나 출력할지를 결정한다[5].

GRU는 2014년 발표된, LSTM의 구조를 보다 단순하게 처리한 LSTM 변형모델의 하나이다[6]. GRU는 LSTM의 망각게이트와 입력게이트를 갱신게이트로 통합하고, 셀 상태와 은닉 상태를 하나로 통합하였다. LSTM보다 단순한 구조로 가중치 수가 작으므로 학습이 더 빠르지만, LSTM과 거의 동일한 성능을 보인다[7]. 본 논문에서 CNN과 결합하여 같이 사용할 모델로 GRU를 선택한 이유이다.

Ⅲ. 본론

본 논문에서 음성 데이터를 분석하고 감정을 분류하기 위해 음성 데이터를 MFCC(Mel-Frequency Cepstral Coefficients)로 전처리하고 학습 및 테스트 데이터로 정규화하는 과정을 CNN-GRU모델 구현 전에 수행한다. 이를 위해 librosa 라이브러리를 사용하여 음성 데이터를 MFCC로 변환시킨다. MFCC로 변환된 음성 데이터는 시간에 따른 주파수 영역의 특징을 포함하고있다. 따라서 CNN에 입력 될 수 있게 2D 형태의 데이터로 변환한다. 이후 TensorFlow를 사용하여 Sequential 모델을 생성한다. Sequential 모델은 레이어를 순차적으로 연결하여 모델을 구성하는 데 사용된다.

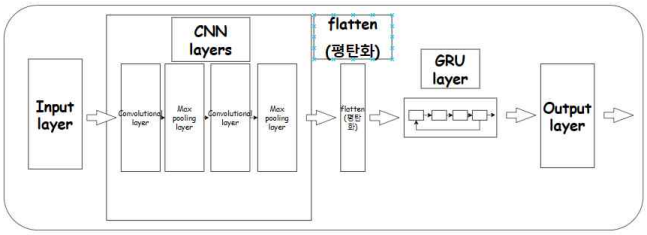

그림 2. CNN-GRU Architecture

<그림 2> 첫 번째 레이어로는 CNN(Convolutional Neural Network) 레이어를 추가한다. CNN은 공간적인 특징을 추출하는 데 효과적이다. 따라서 MFCC로 변환된 음성 데이터의 공간적인 특징을 추출하기 위해 CNN 레이어를 추가한다. 두 번째 레이어는 CNN 레이어의 출력은 3D 형태이기 때문에, 이를 GRU레이어로 보내기 위해 2D 형태로 평탄화(flatten)한다. 이는 데이터의 모양을 변경하여 시퀀스 형태의 데이터를 생성하는 과정이다. 세 번째 레이어는 GRU(Gated Recurrent Unit) 레이어를 추가하여 시간적인 특성을 고려한다. GRU는 시계열 데이터를 처리하는 데 사용된다. 마지막으로, 출력 레이어를 추가하여 감정을 분류한다. 출력 레이어는 분류 작업에 사용되며, 감정 분류 작업을 위해 활성화 함수를 적용하여 각 클래스에 대한 확률 값을 얻을 수 있다. 이를 통해 입력된 음성 데이터의 감정을 분석하는 CNN-GRU모델을 생성한다.

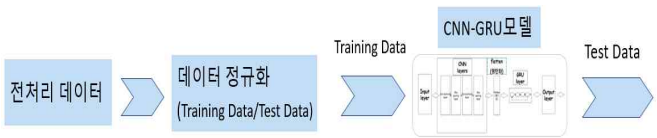

그림 3. CNN-GRU모델 학습 및 테스트

그리고 <그림 3>와 같이 데이터 전처리 및 데이터 정규화를 거친 학습데이터로 학습시킨 후 테스트 데이터로 그 모델의 성능을 평가하고자 한다.

본 논문에서 사용하고자 하는 음성 데이터 셋은 KAIST 인공지능연구소에서 구축된 대화 어플리케이션을 이용하여 수집된 7가지 감정(happiness, angry, disgust, fear, neutral, sadness, surprise)으로 라벨링된 10012개의 음성 데이터 셋을 사용하고자 한다. 10012개의 음성 데이터 파일을 MFCC 형태로 변환해준 뒤 데이터 셔플링(데이터셋을 무작위로 섞어 순서에 따른 편향을 방지)을 해주고 층화 분할(StratifiedShuffleSplit)을 사용하여 데이터를 학습 데이터와 테스트 데이터 80:20비율로 각각의 파일 경로를 저장해주어 8010개의 학습 데이터와 2002개의 테스트 데이터로 나누어 사용하고자 한다.

Ⅳ. 결론

본 논문에서는 사람들의 대화 중 상대방의 감정을 인식하고 이에 맞게 대응하고자, 음성 데이터를 통한 정확한 감정을 인식하여 사회적 상황을 올바르게 인식하는 데 도움이 되고자 음성데이터로 감정을 분석하기 위하여 CNN-GRU을 이용한 모델을 제안하였다. 음성 데이터를 MFCC로 변환하여 시간에 따른 주파수 영역의 특징을 포함한 뒤 CNN을 이용하여 공간적인 특징을 추출한다. 추출된 데이터로 GRU에서 시간적인 특징도 고려하고 마지막 출력 레이어에서 감정을 확률 값으로 출력한다.

CNN-GRU 모델과 기존 연구되었던 CNN-LSTM 모델과의 차이점은 CNN-GRU모델은 한 문장이나 대화의 길이가 긴 장기적 의존성이 크지 않고 짧은 시간 간격 내에 발생하는 음성 신호의 패턴을 가진 데이터를 이용할 때 더 적합하며, LSTM은 셀 상태와 은닉 상태를 별도로 유지하고 조작하는 반면, GRU는 셀 상태와 은닉 상태를 하나로 통합하여 사용하기 때문에 CNN-GRU로 메모리 사용량이 적고 효율적인 모델을 구성할 수 있다. 하지만 대화의 길이가 긴 장기 의존성이 큰 서로 대화하는 형태의 데이터를 이용할 때는 CNN-LSTM이 더 적합하다. 본 연구에서는 서로 대화하는 형식의 데이터가 아닌 혼자 말하는 형식의 데이터를 이용하여 CNN-GRU로 효율적인 감정 분석을 하였다. 이후 서로 대화하는 형식의 데이터에 대해 CNN-GRU를 적용하는 것이 어떻게 되는지 추가적인 연구를 진행하고자 한다.

ACKNOWLEDGMENT

"본 연구는 과학기술정보통신부 및 정보통신기획평가원의 지역지능화혁신인재양성사업의 연구결과로 수행되었음" (IITP-2024-RS-2022-00156287)

참 고 문 헌

[1] Russell, J. A. "A Circumplex Model of Affect Journal of Personality and Social Psychology 39.", Í6I-I78, 1980.

[2] 김주희 and 이석필, "음성 특징과 텍스트 임베딩을 이용한 멀티모달 감정인식," 전기학회논문지, vol. 70, no. 1, pp. 108-113, 2021.

[3] M. Sidorov, S. Ultes and A. Schmitt, "Emotions are a personal thing: Towards speaker-adaptive emotion recognition", Proc. IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), pp. 4803-4807, 2014.

[4] H. Li, X. Zhang and M. -J. Wang, "Research on Speech Emotion Recognition Based on Deep Neural Network," 2021 IEEE 6th International Conference on Signal and Image Processing (ICSIP), Nanjing, China, pp. 795-799, 2021.

[5] 김호현. "LSTM/GRU 순환신경망을 이용한 시계열데이터 예측." 국내석사학위논문 한국방송통신대학교, 2017. 서울

[6] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation", arXiv preprint arXiv:1406.1078, 2014.

[7] Tamal Datta Chaudhuri and Indranil Ghosh, "Artificial neural network and time series modeling based approach to forecasting the exchange rate in a multivariate framework", arXiv preprint arXiv:1607.02093, 2016.