AI 스피커를 활용한 어텐션 메커니즘 기반 멀티모달 우울증 감지 시스템

박준희, 문남미
호서대학교
cach456@gmail.com, nammee.moon@gmail.com

Multimodal depression detection system based on attention mechanism using AI speaker

Junhee Park Nammee Moon
Hoseo University

요 약

전세계적으로 우울증은 정신 건강 질환으로써 문제가 되고 있으며, 이를 해결하기 위해 일상생활에서의 우울증 탐지에 대한 연구가 진행되고 있다. 따라서 본 논문에서는 일상생활에 밀접하게 연관되어 있는 AI 스피커를 사용한 어텐션 메커니즘(Attention Mechanism) 기반 멀티모달 우울증 감지 시스템을 제안한다. 제안된 방법은 AI 스피커로부터 수집할 수 있는 음성 및 텍스트 데이터를 수집하고 CNN(Convolutional Neural Network)과 BiLSTM(Bidirectional Long Short-Term Memory Network)를 통해 각 데이터에서의 학습을 진행한다. 학습 과정에서 Self-Attention 을 적용하여 특징 벡터에 추가적인 가중치를 부여하는 어텐션 메커니즘을 사용한다. 최종적으로 음성 및 텍스트 데이터에서 어텐션 가중치가 추가된 특징들을 합하여 SoftMax 를 통해 우울증 점수를 예측한다.

1. 서론

최근 우울증 장애를 겪는 환자의 수가 매년 증가하면서 전 세계적으로 문제가 되고 있다[1]. 우울증 장애는 발병률에 비해 진단이 어려우며, 제대로 된 치료가 이루어지지 않는 경우에는 자살로 이어지기도 한다.

현재 우울증을 판별하기 위한 방법은 정신 건강 설문지에 의존하고 있으며, 질문에 대한 환자의 반응과 함께 환자 건강 설문지(PHQ, Patient Health Questionnaire), 해밀턴 우울증 척도(HDRS, Hamilton Depression Rating) 또는 벡 우울증 척도(BDI, Beck Depression Inventory) 등이 사용되고 있다[2]. 그러나 BDI 와 HDRS 등 대부분의 척도는 항목수가 많아 실신하는데 시간이 오래 걸리고, 현대적 우울증상과 다른 측면이 많다는 한계가 있다[3]. 따라서 본 논문에서는 우울증을 감지하기 위한 지표로 PHQ 를 사용한다.

또한, 이러한 우울증 문제를 해결하기 위한 방안으로 감성 분석을 통한 우울증 판별 연구가 진행되고 있다[4]. 감성 분석은 감성을 추출하고자 하는 객체로부터 음성, 텍스트 그리고 영상 데이터를 추출하고, 이를 기반으로 감성을 분류한다. 또한, 최근 단일 데이터만을 사용하지 않고 다중 데이터를 사용한 멀티모달 감성 분석을 통해 정확도 향상에 대한 연구가 꾸준히 진행되어지고 있다.

따라서 본 논문에서는 최근 일상생활에서 가까이 접할 수 있는 AI 스피커로부터 음성 데이터와 텍스트 데이터를 수집하여

이를 기반으로 하는 멀티모달 우울증 감지 시스템을 제안한다. 제안된 시스템은 AI 스피커로부터 음성 및 텍스트 데이터를 수집하여 CNN 과 BiLSTM 을 통해 각각의 특징을 추출한다. 그리고 BiLSTM 을 통해 특징을 추출하는 과정에서 높은 가중치를 가지는 어텐션을 추출하여 도출된 특징에 추가로 적용한다. 최종적으로 음성 및 텍스트 데이터에서 각각 도출된 특징과 어텐션을 기반으로 우울증 점수를 예측한다.

2. 관련연구

본 논문의 관련 연구는 우울증 탐지를 위한 기술과 어텐션 메커니즘으로 구성되어 있다.

2.1 우울증 감지

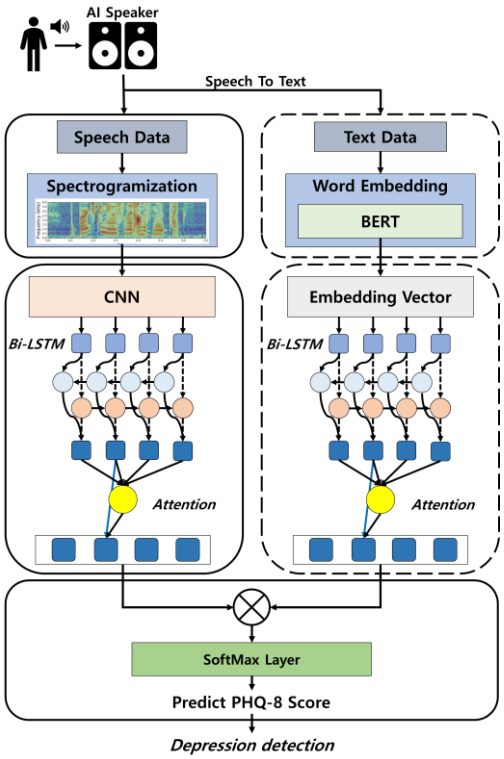
우울증 감지에 대한 연구는 최근 꾸준히 증가하고 있다. Cohn et al.는 사람의 안면과 음성 데이터를 기반으로 기계 학습 분류기를 통해 높은 정확도의 우울증 심각도를 예측하였다[5]. 이후로 Cohn et al.의 실험 결과를 바탕으로 사람의 음성, 영상 그리고 텍스트 데이터를 기반으로 하는 멀티모달 우울증 탐지 모델 연구가 꾸준히 진행되고 있다. Yang et al.은 DCNN(Deep Convolutional Neural Network)와 DNN(Deep Neural Network)을 통해 음성, 영상 및 텍스트를 사용한 멀티모달 우울증 탐지 모델을 제시하였다[6]. Lam at el.은 텍스트 데이터에 사전훈련을 적용한 Transformer 를 적용하고, 오디오의 특징을 추출하기 위해 1 Dimension DCNN 을 사용하여 우울증 감지 모델의 성능을 향상시켰다[7]. Rodrigues et al. 은 멀티모달 우울증 탐지 모델들의 전반적인 성능 비교를 통해 성능의 우수성과 견고함을 증명하였으며, GCNN(Gated Convolutional Neural Network) 과 BERT(Bidirectional Encoder Representations from Transformers) 알고리즘을 사용하여 제안한 모델의 정확도를 개선하였다[8]. Lin at el.은 음성 데이터를 분석하는데 있어 1D CNN 과 Bi-LSTM 모델에서 Mel-Spectrum 변환이 유용한 결과로 도출되었다는 것을 증명하였다[9]. 따라서 본 논문에서는 텍스트 데이터에 높은 정확도 성능을 보이는 BERT 알고리즘 기반 BiLSTM 을 사용하고, 음성 데이터의 특징을 보존하는 MFCC 기반 CNN-BiLSTM 모델을 사용하는 멀티모달 우울증 감지 시스템을 제안한다.

2.2 어텐션 메커니즘

어텐션 메커니즘은 RNN(Recurrent Neural Network)에 기반한 Seq2Seq (Sequence-to-Sequence)모델의 문제점인 정보손실의 문제와 기울기 소실 문제를 해결하기 위한 방법이다. 은닉층의 가중치를 기반으로 특정 벡터에 추가적인 가중치를 높이는 방법으로 최근에는 우울증 탐지와 감성분석에서 이를 활용한 연구가 진행되고 있다. Basiri et al.은 CNN-BiLSTM 에 어텐션 메커니즘을 추가하여 데이터 길이에 상관없는 감성 분석 결과를 도출하였다[10]. Ray at el.은 우울증 예측을 위해 계층적 모델과 어텐션 기법을 사용하였으며, 데이터의 양방향적 특성을 고려한 Bi-LSTM 으로 기존 모델보다 높은 성능을 보였다[4]. Adria at el. 은 NHN(Naıve Hierarchical Network), HLGAN(Hierarchical Local Global Attention Network) 그리고 HCAN(Hierarchical Contextual Attention Network)로 이루어진 계층적 모델 기반 우울증 감지 모델을 제안하였다. 기계학습 모델에서 어텐션 가중치를 부여하는 방법이 우울증 감지 분야에서 적합함을 검증하였다[11]. 본 논문에서는 어텐션 메커니즘을 통해 우울증에 대한 특징과 그렇지 않은 특징에 대한 가중치를 부각시킴으로써 높은 정확도의 우울증 판별을 위한 도구로 사용한다.

3. 시스템 개요

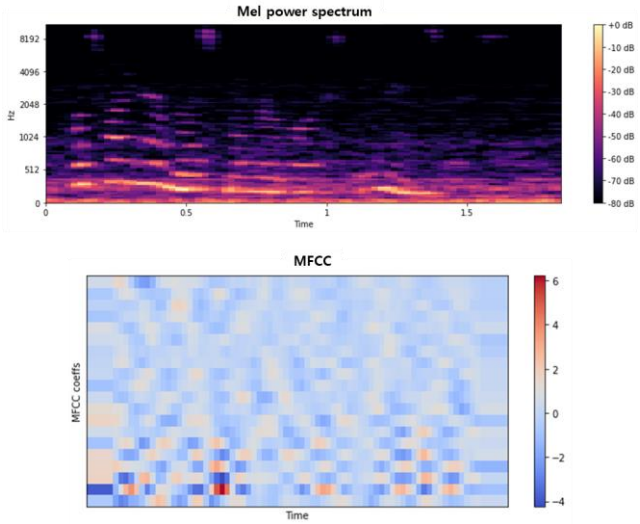
(그림 1)은 본 논문에서 제안하는 시스템의 전체적 시스템 개요이다. 제안된 시스템은 AI 스피커를 통해 데이터를 수집한다. 수집된 서로 다른 양식의 데이터 타입을 각각의 전처리를 통해 임베딩하고, 어텐션 메커니즘이 포함된 2D CNN 과 BiLSTM 모델을 사용해 입력 데이터의 특징을 가중하여 학습한다. 최종적으로 각각의 데이터에서 학습된 정보를 연결하여 SoftMax Layer 를 통해 우울증 점수를 예측한다.



(그림 1) 시스템 개요

3.1 음성데이터 특징 추출

음성 데이터를 CNN 모델에서 학습시키기 위해서는 데이터를 이미지화 시키는 사전 작업을 필요로 한다. 따라서 시간에 따른 음압 표현으로 이루어진 음성 데이터를 주파수 영역의 표현으로 바꾸어 주기 위해 FFT(Fast Fourier Transform)을 적용하여 Spectrum 으로 변환한다. Spectrum 은 물리적인 주파수와 실제사람이 인식하는 주파수의 관계 형태를 표현한 Mel scale 기반 Filter Blank 를 적용하여 (그림 2)의 Mel-Spectrum 을 구한다. 최종적으로 Mel-Spectrum 은 소리의 고유한 특징을 찾아주는 Cepstral 분석을 사용하여 모델에 사용하기 위한 MFCC(Mel-Frequency Cepstral coefficient)로 변환한다. 도출된 MFCC 는 2D CNN 알고리즘을 통해 특징 벡터를 추출하고 Bi-LSTM 을 통해 학습을 한다. 학습과정 중 은닉층에서 각 셀의 정보를 기반으로 우울증에 대한 특징 정보인 어텐션을 수집한다. CNN 을 통해 도출된 특징 정보와 특징정보에 부가적인 가중치를 주는 어텐션을 BiLSTM 모델에 적용하는 Self-Attention 방법을 사용한다.



(그림 2) Mel-Spectrum 및 MFCC

3.2 텍스트 데이터 특징 추출

AI 스피커를 통해 수집되는 데이터는 음성 데이터이므로 텍스트 데이터를 수집하기 위해서 입력된 음성 데이터는 STT(Speech-To-Text)를 이용하여 텍스트 데이터로 변환한다. 텍스트 데이터는 BERT 알고리즘을 적용하기 위해 정규화 및 전처리를 하는 NLTK 모듈을 사용한다. WordNetLemmatizer 를 통해 표제어와 어간을 추출하고, 대소문자 통합 작업, 불용어(Stopword)를 제거한다. 전처리된 데이터는 BERT 알고리즘을 사용하여 임베딩 벡터를 추출하고, 이를 BiLSTM 모델을 통해 학습을 진행한다. 또한 학습과정에서 특징을 부각시키기 위해 Self-Attention 을 사용한다.

3.3 우울증 점수 예측

두 개의 모델에서 나온 어텐션이 적용된 텍스트 데이터와 음성 데이터를 Fully Connected Layer 에서 데이터를 융합한다. 최종적으로 두 데이터가 융합된 임베딩 벡터를 SoftMax Layer 를 사용하여 PHQ-8 스코어 예측을 통해 우울증을 감지한다.

4. 결론

본 논문은 사회적 이슈가 되고 있는 우울증을 예측하기 위해 AI 스피커를 통한 어텐션 메커니즘 기반 멀티모달 우울증 예측 시스템을 제안하였다. 제안된 모델은 다양한 데이터를 사용한 멀티모달 데이터 분석을 통해 높은 정확도를 기대할 수 있을 것이다.

또한, 어텐션 메커니즘과 CNN 및 BiLSTM 모델의 학습 능력을 통해 다양한 형식의 데이터에서 임베딩 벡터를 도출하는데 좋은 효과가 있을 것이다.

향후 연구에서는 본 논문에서 제안된 시스템에 대한 실험과 음성과 텍스트 데이터 외의 추가적인 데이터를 사용한 우울증 예측 시스템을 진행할 것이다.

사사

본 연구는 과학기술정보통신부와 정보통신기획평가원의 SW 중심대학사업의 연구결과로 수행되었음 (2019-0-01834)

참고문헌

[1] Hassan, A. U., Hussain, J., Hussain, M., Sadiq, M., Lee, S., “Sentiment analysis of social networking sites (SNS) data using machine learning approach for the measurement of depression.” In 2017 International Conference on Information and Communication Technology Convergence (ICTC), IEEE, 2017, p. 138-140.

[2] Park, S. J., Choi, H. R., Choi, J. H., Kim, K. W., Hong, J. P., “Reliability and validity of the Korean version of the Patient Health Questionnaire-9 (PHQ-9).” Anxiety and mood, 2010, 6(2), p. 119-124.

[3] Je Yong An, Eun Ran Seo, Kyung Hi Lim, Jae Hyun Shin, Jung Bum Kim., “Standardization of the Korean version of Screening Tool for Depression(Patient Health Questionnaire-9” PHQ-9). JOURNAL OF THE KOREAN SOCIETY OF BIOLOGICAL THERAPIES IN PSYCHIATRY, 2013, 19(1), p. 47-56.

[4] Ray, A., Kumar, S., Reddy, R., Mukherjee, P., Garg, R., “Multi-level attention network using text, audio and video for

depression prediction.” In Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop, 2019, p. 81-88.

[5] Cohn, J. F., Kruez, T. S., Matthews, I., Yang, Y., Nguyen, M. H., Padilla, M. T., ... & De la Torre, F. (2009, September). Detecting depression from facial actions and vocal prosody. In 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, IEEE, 2009, p. 1-7.

[6] Yang, L., Jiang, D., Han, W., Sahli, H., “DCNN and DNN based multi-modal depression recognition.” In 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), IEEE, 2017, p. 484-489.

[7] Lam, G., Dongyan, H., Lin, W., “Context-aware deep learning for multi-modal depression detection.” In ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2019, p. 3946-3950.

[8] Rodrigues Makiuchi, M., Warnita, T., Uto, K., Shinoda, K., “Multimodal fusion of BERT-CNN and gated CNN representations for depression detection.” In Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop, 2019, p. 55-63.

[9] Lin, L., Chen, X., Shen, Y., & Zhang, L., “Towards Automatic Depression Detection: A BiLSTM/1D CNN-Based Model.” Applied Sciences, 2020, 10(23), p. 8701.

[10] Basiri, M. E., Nemati, S., Abdar, M., Cambria, E., Acharya, U. R., “ABCDM: An attention-based bidirectional CNN-RNN deep model for sentiment analysis.” Future Generation Computer Systems, 2021, 115, p. 279-294.

[11] Mallol-Ragolta, A., Zhao, Z., Stappen, L., Cummins, N., Schuller, B., “A hierarchical attention network-based approach for depression detection from transcribed clinical interviews”, 2019.