앙상블 기법의 도시 사운드 분류 모델 성능 비교 분석

A Comparative Analysis on Ensemble technique of Urban sound classification model

Gyeongmin Kim[1], YunA Hur[1], Aram So[1], Heuiseok Lim[2*]
[1] *Integrated Master's and Ph.D. course, Dept of Computer Science and Engineering, Korea University,145 Anam-ro, Seongbuk-gu, Seoul, 02841, Korea*
[2*] *Professor, Dept of Computer Science and Engineering, Korea University,145 Anam-ro, Seongbuk-gu, Seoul, 02841, Korea*
totoro4007@korea.ac.kr[1], yj72722@korea.ac.kr[1], aram@korea.ac.kr[1], limhseok@korea.ac.kr[2]*
Corresponding author: Heuiseok Lim*

Abstract Recently, in the advancement of deep learning-based sound classification modeling, a lot of studies using various neural network techniques in sound have been modeled. In this paper, we evaluate the performance of the sound classification model based on deep learning models and analyze them with ensemble results. The results of performance showed that 90.40% in our CNN model and 93.25%, which is 2.85 points higher than CNN in the ensemble results.

Keywords: *sound classification, ensemble, deep learning*

1. 서론

최근 음성합성, 화자분리, 화자인식과 같은 음성인식 전반적인 기반 기술의 발전과 이를 응용한 대화형 키오스크 시스템 및 챗봇, 무인상담 시스템, 실시간 강연 통역 등과 같은 음성, 소리와 접목한 딥러닝 기술이 등장하고 있다. 음성인식 기술은 방송, 전화상담, 녹취, 통역, 빅데이터 분석 등의 시장에서 음성데이터 분석에 대한 그 수요가 커지고 있으며, 최근 음성인식 분야에서는 noise 를 분리하려는 다양한 연구가 진행되고 있다[1-3]. 대표적인 예시로 Urban Sound Classification 데이터셋을 이용하여 도시의 다양한 소리에 대해 모델링하는 연구가 증가하고 있고, 가공되지 않은 음성데이터로부터 noise 를 제거하여 음성인식의 성능을 높이려는 연구가 등장하고 있다[4,5].

해당 기술은 화자 인식 기술을 변형한 기술이라고 정의할 수 있다. 화자인식이란 화자인증 기술과 화자식별 기술로 나뉘며 화자 인증은 문장 독립 화자인증과 문장 종속 화자인증으로 나뉜다. 화자인식(Speaker recognition)은 음성이 입력으로 들어왔을 때 그 발화 내용이 아닌 발화자를 출력으로 내보내는 기술이다. 화자 식별(Speaker identification)이란 등록된 N 명 중 가장 유사한 1 명을 찾는 기술을 의미하며 화자 인식 기술의 하위기술이다. 즉 해당 기술은 음성에서 특정 부분을 추출하여 분리하는 기술로 정의할 수 있으며, 본 논문에서 다루는 소리 분류 기술은 이와 유사한 특징을 지니고 있어 각종 챗봇 시스템, 무인 상담 시스템의 응답률을 높이는 각 서비스에 활용 가능하다. 또한, 비화자 목소리로 인한 명령어 처리 방해를 해소할 수 있으며 이로 인해 서비스 응답률을 상승시킬 수 있다. 즉 소리 분류 모델은 실질적으로 다양한 어플리케이션에 적용 가능한 기술이다.

본 논문은 이러한 필요성을 바탕으로 기 구축된 도시 소리 분류 데이터셋으로부터 자질(features)을 추출하여 다양한 딥러닝 모델로 학습시키고, 각각의 모델에 대한 성능 평가와 앙상블 모델의 성능 향상 여부를 비교 분석한다.

2. 본론

본 논문의 연구 순서는 학습 데이터 수집, 전처리, 자질 생성, 모델 구성, 모델 학습 및 평가로 진행한다.

Urban sound classification 데이터셋은 도시에서 발생 가능한 소리데이터 10 가지 종류의 소리데이터로, 전체 5547 개의 데이터를 학습 및 평가 과정에 사용한다. 소리데이터는 모델 학습을 위해 기존 아날로그 형태의 데이터를 디지털 신호로 변환시키고, 그 데이터로부터 다양한 디지털 신호 자질을 추출하여 학습에 사용한다. 대표적으로 MFCCs(Mel-Frequency Cepstral Coefficients, MFCCs)와 같은 음성 자질을 모델 학습을 위한 자질로 사용할 수 있고 본 연구에서 사용한 음성 자질은 MFCCs 외에도 Mel-spectrogram, chroma_stft, chroma_cqt, chroma_cen 를 학습 자질로 사용한다.

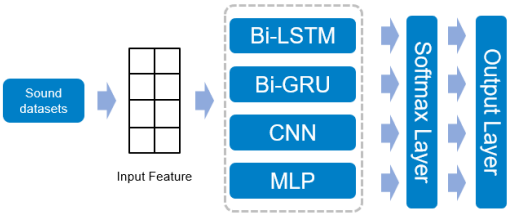

Figure 1. Training process for each model

순환신경망(Recurrent Neural Network, RNN) 모델과 합성곱 신경망(Convolutional Neural Network, CNN) 모델은 이미지, 텍스트뿐 만이 아니라 음성인식 자질을 활용한 모델 학습에 활용할 수 있으며, 본 연구에서는 위 [Figure 1.]와 같이 RNN 기반의 Bi-LSTM(Bi-directional Long Short Term Memory, Bi-LSTM), Bi-GRU(Bi-directional Gated Recurrent Unit, Bi-GRU) 모델과 CNN, 그리고 MLP(Multi-Layer Perceptron, MLP) 모델에 대해 성능을 측정한다. 그리고 위 4 가지 모델을 앙상블 실험 결과에 대한 성능 향상을 비교 분석한다.

4. 실험결과

Table 1. Performance comparison analysis

Model	***Epoch***	***Validation set Accuracy***	***Test set Accuracy***
MLP	150	90.45%	89.65%
CNN	70	91.02%	90.40%
Bi-GRU	50	86.76%	86.24%
Bi-LSTM	40	88.45%	88.26%
Ensemble			**93.25%**

평가를 위해 모델의 정확도(accuracy)를 측정하였으며 학습 결과는 [Table 1.]과 같다. 모든 모델에서 test set 보다 validation set 에서 0.5% 높은 정확도를 보여주었고, MLP, CNN 모델에서 문자열과 같은 순차적 데이터 처리 RNN 기반 모델보다 좋은 성능을 보여준 것을 확인할 수 있었다. 위 4 개의 모델을 앙상블한 결과 93.25%의 성능으로 가장 좋은 성능을 보여준 CNN 모델 대비 약 2.85% 성능 향상이 있었던 것을 확인할 수 있었다.

5. 결론

본 논문은 도시 사운드 분류 데이터를 활용하여 다양한 딥러닝 모델에 적용시켜봄으로써 분류 모델 성능을 평가한다. CNN 모델에서 가장 좋은 성능을 보였으며, 앙상블 결과 CNN 보다 약 2.85% 좋은 성능을 보여준 것을 확인할 수 있었다.

Acknowledgments

This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2020-2018-0-01405) supervised by the IITP(Institute for Information & Communications Technology Planning & Evaluation) and National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP) (No.NRF-2017M3C4A7068189)

References

[1] J. G. Lee., & B. H. Lee. (2019). Motion study of Treatment Robot for Autistic Children Using Speech Data Classification Based on Artificial Neural Network, Journal of IKEEE, 23(4), 325-332.

[2] H. G. Kim., G. J. Jang., & H. J. Choi. (2019). Pitch Classification Based on Bidirectional LSTM with Probabilistic Attention for Speech Segregation from Speech-Music Mixtures. Proceedings of Symposium of the Korean Institute of communications and Information Sciences, KIISE Transactions on Computing Practices, 25(4), 223-230.

[3] AbuShawar, B., & Atwell, E. (2016). Usefulness, localizability, humanness, and language-benefit: additional evaluation criteria for natural language dialogue systems. International Journal of Speech Technology , 19 (2), 373-383.

[4] D. S. Park., J. I. Bang., & Y. J. Ko. (2018). A Study on the Gender and Age Classification of Speech Data Using CNN, Journal of KIIT, 16(11), 11-21.

[5] Y. J. Lee & J. H. Chang (2017). A Study of the Convolutional Neural Network Structure for Urban Sound Classification. Proceedings of Symposium of the Korean Institute of communications and Information Sciences, 957-958.