Classification of Musical Instruments using Convolutional Neural Network

Sun Woo Park, Byeong Jun Park, Sangtae Ahn*
School of Electronics Engineering, Kyungpook National University
E-mail: psw9808@naver.com, qudwns7171@naver.com, stahn@knu.ac.kr

Method:
In the field of deep learning, many studies on sound data processing have been conducted. Most of the studies were natural language processing [1] and text-to-speech [2], but there were few studies on the problem of Musical Instrument Classification (MIC) [3].
Spectrogram is one of the methods for processing sound data. It is a tool for visualizing and grasping sounds or waves and a combination of waveform and spectral features. Through the spectrogram, the sound classification problem can be applied as an image classification problem. The convolutional neural network (CNN) is used to process image data and is well applied to image recognition and classification problems. In this work, thus, we aim to designed a deep learning model using CNN to solve the MIC problem. The figure below shows the model structure we designed.

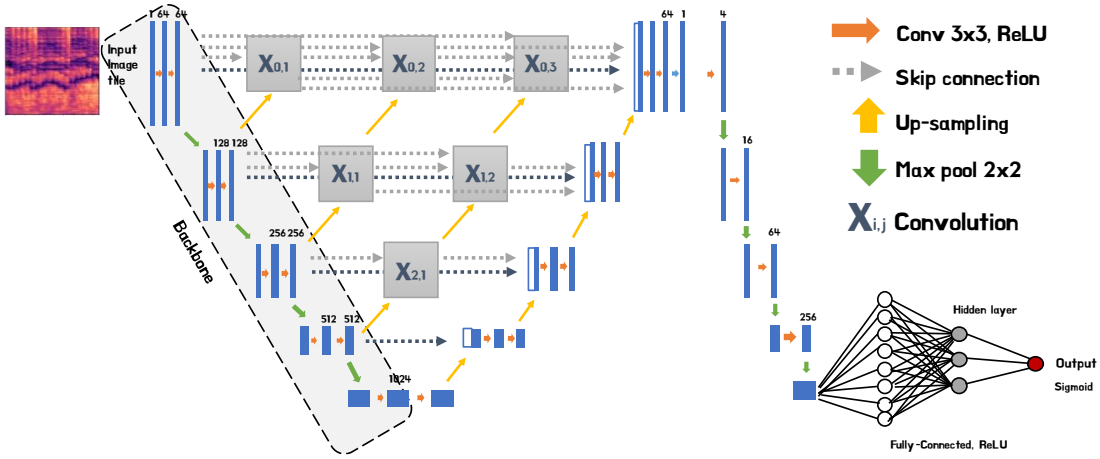

Figure 1. Our suggested model architecture

U-Net [4] is an encoder/decoder CNN architecture with skip connections, which has been used for vocal separation [5]. We have added UNet++ [6], which is a redesigned U-Net that has a deeply-supervised encoder/decoder network. We used multi-label classification [7] to distinguish whether each instrument class is present or not. To this end, each model is designed with a binary classification method. Thus, a set of outputs from each model generated the final output.

Experimental Results:
We used a dataset of 28,668 data collected from 577 videos on the web. The dataset consists of 7 instruments (cello, clarinet, drum, flute, piano, viola, violin). We preprocessed the data into spectrogram using the constant-Q transform (CQT) [8] and pass it through the model.

The table below shows the test accuracy of the model for each musical instrument.

Table 1. Classification accuracy of each musical instrument.

Instrument	Cello	Clarinet	Drum	Flute	Piano	Viola	Violin
Accuracy(%)	87.85	84.90	97.57	84.58	93.53	77.91	84.44

The results show reasonably high classification accuracies for each musical instrument. However, low accuracy for three or more ensembles also appears. We observed that the sound range of the instrument affected the classification performance of the model. We expect improved classification performance if the dataset is well organized. In the view of these features, there is a possibility that this sound classification mechanism may be applied not only to instruments but also to other subjects.

Keyword: spectrogram, UNet++, multi-label classification, constant-Q transform

Reference

[1] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-tau Yih, T. Rocktäschel, S. Riedel, and D. Kiela, "Retrieval-augmented generation for knowledge-intensive NLP tasks," *arXiv.org*, 12-Apr-2021.

[2] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan, R. A. Saurous, Y. Agiomvrgiannakis, and Y. Wu, "Natural TTS synthesis by conditioning wavenet on Mel Spectrogram predictions," *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018.

[3] K. Racharla, V. Kumar, C. B. Jayant, A. Khairkar, and P. Harish, "Predominant musical instrument classification based on spectral features," *2020 7th International Conference on Signal Processing and Integrated Networks (SPIN)*, 2020.

[4] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *Lecture Notes in Computer Science*, pp. 234–241, 2015.

[5] R. Hennequin, A. Khlif, F. Voituret, and M. Moussallam, "Spleeter: A fast and efficient music source separation tool with pre-trained models," *Journal of Open Source Software*, vol. 5, no. 50, p. 2154, 2020.

[6] Z. Zhou, M. M. Rahman Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net Architecture for Medical Image segmentation," *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pp. 3–11, 2018.

[7] J. Read and F. Perez-Cruz, "Deep learning for multi-label classification," *arXiv preprint arXiv:1502.05988*, 2014.

[8] K. W. Cheuk, K. Agres, and D. Herremans, "The impact of audio input representations on neural network based music transcription," *2020 International Joint Conference on Neural Networks (IJCNN)*, 2020.