Musical Instrument Classification using Convolutional Neural Network

Sun Woo Park and Byeong Jun Park

School of Electronic and Electrical Engineering, Kyungpook National University

E-mail: [psw9808@naver.com](mailto:psw9808@naver.com) , [qudwns7171@naver.com](mailto:younghe@ee.knu.ac.kr)

**Method:**

In the field of machine learning and deep learning, many research on sound data processing is being conducted. Among them, there are many studies such as Natural Language Processing (NLP)[1], Text-to-Speech (TTS)[2], etc. but there are 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 like images. It is a combination of waveform and spectral features. Through the spectrogram, the sound classification problem can be applied as an image classification problem. Convolutional Neural Network (CNN) is used to process image data and is well applied to image recognition and classification problems.

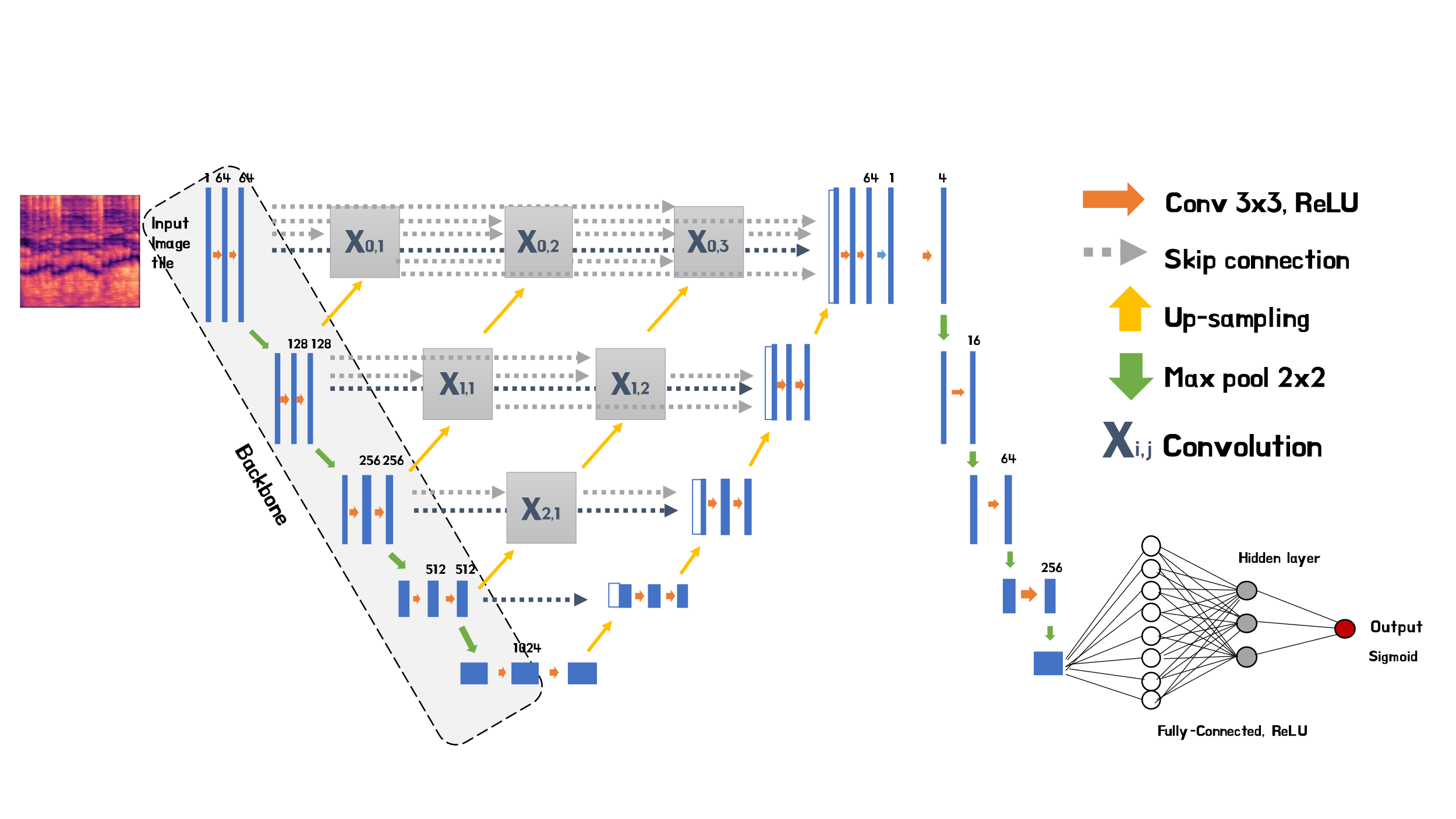
In this paper, we aim to create our own model using CNN to solve the MIC problem. The figure below shows the model structure we designed.

Fig 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 added UNet++[6], which is a redesigned U-Net that has a deeply-supervised encoder/decoder network.

We use 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. So, a set of outputs from each model draws a final conclusion.

**Experimental Results:**

We used a dataset of 28,668 data collected from 577 YouTube videos. The dataset consists of 7 instruments (Cello, Clarinet, Drum, Flute, Piano, Viola, Violin). We preprocess data into spectrogram using Constant-Q Transform (CQT)[8] and pass it through the model. The table below shows the test accuracy of the model for each instrument.

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

Table 1. The classification accuracy by musical instrument of our experiments

The results show high accuracy for a single instrument. However, low accuracy for three or more ensembles also appears. The sound range of the instrument affected the classification performance of the model. Better performance is expected if the dataset is improved. In 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, CQT (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.