CS230

Project Proposal for Automatic Music Transcription
Project Category: Sound Recognition

Ruoyan Chen
Department of Electrical Engineering
Stanford University
ruoyan85@stanford.edu

Yiwen Liu
Department of Electrical Engineering
Stanford University
ywliu24@stanford.edu

1 Problem Description

For a long time, the music sheet has been regarded as one the most effective media for musicians to communicate with each other. It is also an intuitive way for non-professionals to learn how to play a music instrument or sing a song. Nevertheless, music sheet might not be available for all compositions especially for those being protected by strict copyright regulations. Therefore, to better provide music beginners or amateurs with a chance to play these compositions, we came up a project idea of using deep learning techniques to perform automatic music transcription. The input to our algorithm would be raw audios. After training, our model will be able to translate music audios into the symbolic music representation output in the format of musical alphabet.

2 Challenge

After the literature review, we found that there's a common technique used in the data pre-processing which is down-sampling audio signals to reduce the amount of data. This would be helpful for improving the run-time efficiency. For the network architectures, the combination of Convolutional Neural Networks (CNNs) and Recurrent NeuralNetworks (RNNs), as well as some comparisons between Deep Neural Network (DNNs) and LongShort-Term Memory (LSTM) are often mentioned. Inspired by those information, for this project we would like to explore more on developing the network architectures such as comparing performances among various models and make some combination of them to improve the performance.

3 Dataset and Features

For this project, we are going to use NSynth [1], which is an open-source audio database. The full dataset is split into three sets: a training set with 289,205 examples; a validation set with 12,678 examples; and a test set with 4,096 examples. For data pre-processing, we plan to first down-sample the audio signals, and also normalize the extracted features. We are also thinking about applying Q Transform to further improve the run-time efficiency. From the provided features, we will extract pitch and velocity as the label for training.

4 Methods

The most common method to do music transcription is to first downsample audio signal to create its spectrogram. Then use supervised learning based on each note's pitch. The objective is to minimize the loss defined by the difference between the predicted and the real pitch. In existing researches [2]

[3] [4] [5] [6], several network architectures are implemented, including DNNs, RNNs, and CNNs. In our project, we want to explore a combination of networks to achieve better results.

5 Evaluation

We will evaluate the model by its prediction accuracy on pitch and beat separately. The test set will contain both monophonic and polyphonic music pieces. We will show the prediction accuracy changing through training episodes, and will compare the performances of different network architectures. We may also include additional evaluation on our own dataset.

References

[1] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Douglas Eck, Karen Simonyan, and Mohammad Norouzi. Neural audio synthesis of musical notes with wavenet autoencoders, 2017.

[2] Curtis Hawthorne, Erich Elsen, Jialin Song, Adam Roberts, Ian Simon, Colin Raffel, Jesse Engel, Sageev Oore, and Douglas Eck. Onsets and frames: Dual-objective piano transcription. *arXiv preprint arXiv:1710.11153*, 2017.

[3] Yu-Lun Hsu, Chi-Po Lin, Bo-Chen Lin, Hsu-Chan Kuo, Wen-Huang Cheng, and Min-Chun Hu. Deepsheet: A sheet music generator based on deep learning. In *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 285–290. IEEE, 2017.

[4] Jonggwon Park, Kyoyun Choi, Sungwook Jeon, Dokyun Kim, and Jonghun Park. A bi-directional transformer for musical chord recognition. *arXiv preprint arXiv:1907.02698*, 2019.

[5] Miguel A Román, Antonio Pertusa, and Jorge Calvo-Zaragoza. A holistic approach to polyphonic music transcription with neural networks. *arXiv preprint arXiv:1910.12086*, 2019.

[6] Siddharth Sigtia, Emmanouil Benetos, and Simon Dixon. An end-to-end neural network for polyphonic piano music transcription. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(5):927–939, 2016.