Automatic Music Transcription with LSTM and DNN

Gracie Zhou, Qianyu Zhu, Louisa Liu

1 Introduction

Machine learning has demonstrated its potential in solving a variety of problems, ranging from music generation and recommendation, to music visualization and automatic music transcription. Music lovers frequently encounter the following scenario: a beautiful melody whisps their mind before it can be preserved permanently as sheet music, due to a lack of technical training. Is there a way to simplify music creation so as to empower non-musicians? Moreover, given a piece of improvised music, can we efficiently and accurately transform it into music sheets to faciliate revision and reproduction? Inspired by these questions, our team attempts to automatically transcribe music recordings into sheet music.

In the paper *Onsets and Frames: Dual-objective Piano Transcription*, Google Brain Team defines the following objective:"Automatic music transcription (AMT) aims to create a symbolic music representation (e.g., MIDI) from raw audio." [1] Current efforts in this field revolve around deep neural network (DNN) and long short-term memory networks (LSTM), which our team will utilize and make further improvements on.

2 Data Overview

This project uses the MAESTRO dataset from magenta, containing more than 200 hours of piano performances. The fact that there is only one musical instrument present in the recordings greatly faciliates data analysis, ruling out the possibility of track-overlapping. Training data consists of 1282 mp3 recordings with their corresponding MIDI files, and a csv file containing the names of the pieces and their composers, the years of the recordings and their durations. The dataset has already been splitted into training (80%), validation (10%) and testing (10%) sets. To make sure that the algorithm would not cheat, the splits are conducted in such a way that one composition is included in only one of the sets.

3 Methodology and Algorithm

The most popular algorithms are multi-label logistic regression (baseline model), deep neural network (DNN), and long short-term memory network (LSTM) [2]. In the latest published papers, research teams mainly use the method of "Onsets and Frames", which splits note detection tasks across two stacks of neural networks [1]. One stack is trained to detect only onset frames (the first few frames of every note), and the other stack is trained to detect every frame where a note is active. Motivated by their work, our team will build a similar model based on LSTM. We choose tanh (hidden layer) and sigmoid (output layer) as the action functions, and binary cross-entropy (BCE) as the loss function. Considering the inefficiency associated with audio files processing, we will transform the input .mp3 files into spectrograms, using constant Q transform (CQT), in order to extract relevant features [2]. After preprocessing the dataset by downsampling and normalizing, we will get a data matrix with length denoting time frame and width denoting the number of features (in the form of 0/1 vector, 1 represents notes/chords that are present). The expected output of the model is a MIDI file containing the information of every note in the input audio. Stacking, stochastic gradient descent, and Onsets and Frames will be considered for further improvement.

During the final evaluation, we will use accuracy to test our machine learning model's prediction power. We will also utilize an external package Music21 to generate music scores synchronously.

References

[1] 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.*

[2] Luoqi Li, Isabella Ni, Liang Yang. *Music Transcription Using Deep Learning.*