Project Work - final submission

Deep Learning : Professor, Bálint GYIRES-TÓTH
5/11, Takuto Namba

Introduction

Nowadays, Deep learning is applied to a lot of fields, such as face recognition, self-driving cars and also music generation. I was interested in creating the music from the existing one because I like to listen to all kinds of music and I thought it would be fun to let some non-human to play the music.
The automatic music creation is a process of composing short pieces of music with the training model defined by deep learning technology. In this project, I am exploring two approaches, 1D CNN (One-dimensional convolutional network) and LSTM (Long Short Term Memory) to see which models will give the accurate output that sounds similar to the target output of the dataset.

Dataset

MAESTRO (MIDI and Audio Edited for Synchronous TRacks and Organization), we used for this project, is a dataset composed of about 200 hours of virtuosic piano performances captured with fine alignment (~3 ms) between note labels and audio waveforms. Repertoire is mostly classical, including composers from the 17th to early 20th century.

Proposed methods

As the first trial, I used 1D CNN as this training approach is useful for the time-series data like the music. Considering the previous results I got with 1D CNN, I decided to use LSTM for the next approach. LSTM is a type of recurrent neural network capable of learning order dependence in sequence prediction problems. The main difference between CNN and RNN is the ability to process temporal information or data that comes in sequences, such as a sentence. In other words, RNN is much more suitable for remembering the previous results in sequences.

Evaluation methods

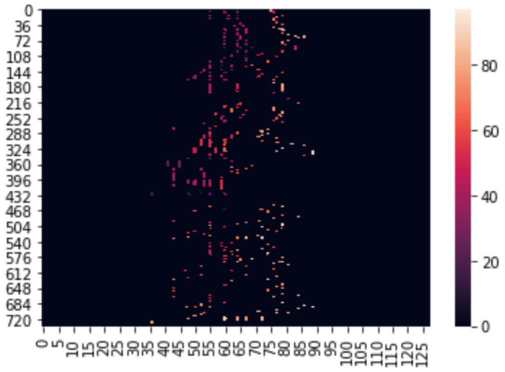
I evaluate the time-series result with a heatmap. A heat map is a graphical representation where individual values of a matrix are represented as colors and is very useful in visualizing the concentration of values between two dimensions of a matrix. Since the music is a collection of different frequencies, the difference of how the frequencies are concentrated in the target music and predicted music is essential.

Results

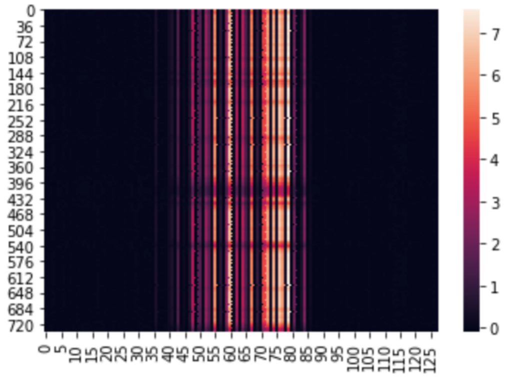
Here are the results in the form of heat map:

First approach with 1D CNN

Target output

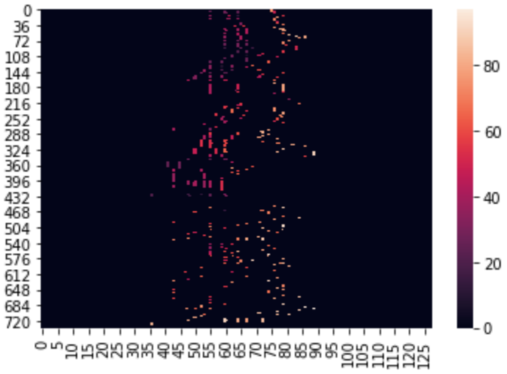


Predicted output

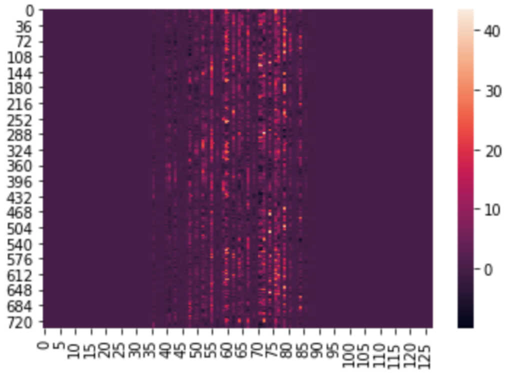


Second approach with LSTM

Target output



Predicted Output



Observation/Discussion

Observation

- In 1D CNN, it only gives the time-series output for one pitch, the sound produced by a single key. This is the reason why the heatmap shows that there are just lines However, this midi file is composed of a lot of nodes. It tells that 1D CNN is not quite useful in terms of creating music with multiple pitches.
- On the other way, the predicted output done with LSTM is quite similar to the target output. Their midi files are also reasonably alike.

Discussion

- As shown in the results, we can deduce that LSTM is much useful to create music with multiple pitches, like most of the music.
- Both target and predicted results do not sound like rhythmic music although they are similar to each other. For the next time, I might need to add more LSTM layers or adapt the Dropout technique in order to reduce more overfitting and give improvements.
- We can try fine-tune a pre-trained model to build a robust system, as the size of the training dataset is small
- The use of WaveNet, a deep neural network for generating raw audio, for the further training might make the better result.