MaestroNet

Musical Transcription Through Seq-to-seq Modelling

Nora Cai, Victoria Chen, Siddharta Laloux, Atharva Nihalani

BROWN Computer Science

Introduction

Transcribing music is a labor intensive process often requiring professional training. We believe that Deep Learning will help non-professionals transcribe and play their favorite songs. Using seq-to-seq modelling, our model converts a recording of piano music to a sheet music representation. This relationship was modeled between recordings in a .wav format and a sheet representation in MIDI, as many research teams in the field had done.

The high performance of off-the-shelf transformers for seq-to-seq tasks make them very suitable for this problem, and hence, that's the architecture we implemented. Our problem statement is thus formulated as follows: **how do we train a transformer to convert .wav music files to an equivalent MIDI file?**

Data

The Saarland Piano Dataset contains 50 paired .mp3 and .midi files of solo piano performances. We decided to split the dataset into 80/10/10 percent for our train/validation/test data respectively. For data preprocessing, we computed a constant-Q spectrogram from the .wav files (Fig. 1).

Using PrettyMidi, we converted the .midi files into binary pianorolls, where each timestep contains either a 0 or 1 to represent the absence or presence of a note (Fig. 1). Since all the performances in the dataset were of varying length, we chose to standardize the inputs and labels by splitting the performances into 5-second segments.

Methodology

Our architecture is a standard transformer-based seq-to-seq model. The model takes in spectrograms of .wav files as input, and outputs a piano roll estimation of note pitch and duration for each time frame. Since transformers can store context indefinitely and choose which parts of its inputs to pay attention to, they are particularly helpful for dealing with classical music, which often contains rich structure and recurring motifs. We also believe the transformer memory would help us model the well-attested forms (e.g. sonata, fugue, and toccata) of classical music.

As transformers are very training-intensive, we decided to use a small model of 3 encoder and 3 decoder blocks, each block containing 3 attention heads, with key and value dimensions of 64. Spectrograms were embedded into 512-element vectors, and the model's feedforward layers output 1024-element vectors.

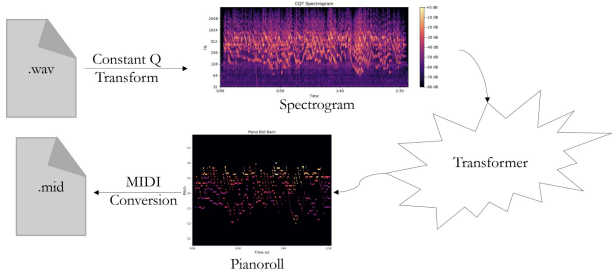

Figure 1. Model architecture and data pipeline

Results

Our best accuracy (86.45%) on held-out test data was achieved after 4 epochs. Average loss on the same test data was 2.54.

This accuracy aligns with our stretch goals, and is very close to MT3's performance on note pitch and length identification.

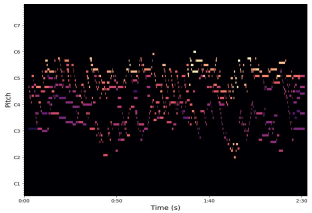

Figure 2. Model output pianoroll for the spectrogram in fig. 1

Discussion

Analysis: We chose pianorolls as our output format for a few reasons. Although as music representations pianorolls are sparser and more memory-intensive, they can be easily generated from MIDI files, and off-the-shelf loss functions can be used. Meanwhile, pianorolls only consider the pitch and duration of a note, which allows a simple architecture to perform well.

Limitations: Our model is unable to predict dynamics due to the simplicity of the pianoroll representation and the complexity of our model. Our model also struggles to transcribe grace notes, possibly due to our time-scale of 10ms or the spectrogram transformation.

Future Directions: We aim to enhance our model to handle music with multiple instruments. Transcribing additional instruments necessitates capturing dynamic changes (such as crescendos and diminuendos) and pitch variations within a note (like glissandos). Thus, we plan to develop a more comprehensive event vocabulary capable of describing dynamics.