Midi Shark: For Piano Transcription

Jonah Chen, QiLin Xue, Joe Hattori, Khanatat Thangwatthanarat
University of Toronto
{jonah.chen,qilin.xue,joe.hattori,k.thangwatthanarat}@mail.utoronto.ca
December 8, 2021

1 Introduction

Transcription of music is the process of determining the pitches and timing of notes from recorded audio files. Transcription has always been a specialized task that requires years of musical training. Transcription is even more challenging for polyphonic music, such as piano, which features the simultaneous production of two or more tones. The majority of traditional transcription models focus on extracting all of the notes from the recording using the node onset. This, however, is not the way a trained musician approaches the problem[1].

We developed a model that is more accurate at transcribing piano recordings by analyzing the recording with a neural network and focusing on both the onsets and offsets of the node. Moreover, since images and audios both have common two-dimensional time-frequency input representations, the fact that CNN performs well in image classification problems suggests that CNN could potentially be used for music transcription[2].

2 Illustration

The model architecture is shown in figure 1.

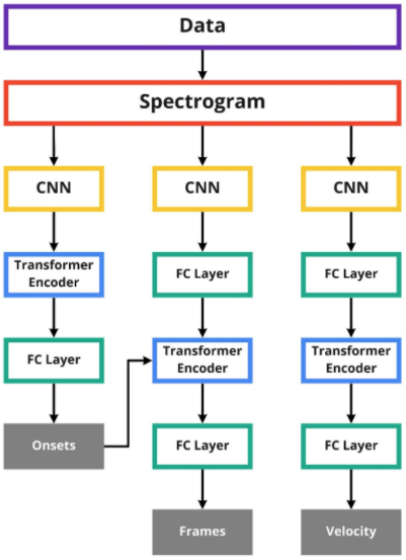

Figure 1: Model architecture

3 Background

Polyphonic automatic music transcription is a difficult task due to the potential of having multiple notes played simultaneously, causing the harmonics of each note to overlap with the others. Therefore, it is a non-trivial task to transcribe the music using the spectrogram.

Since the advent of machine learning, there have been models developed for solving this problem. A relatively successful model is the Onsets and Frames (OF) model developed by the Google Brain Team using TensorFlow in 2017 and implemented into the open-source "magenta" framework[9]. This model uses a two-headed encoder-decoder architecture that predicts onsets (when the notes are played), and frames (what notes are played at the onset). This model uses convolutional neural networks as encoders and bidirectional LSTM as decoders and was able to achieve an F1 score of 78.30% and 82.29% on frames and notes respectively for the MASTERO dataset, which is arguably the largest dataset for music transcription task.

4 Data Processing

We used Maestro Dataset as the Google Brain Team did to develop the OF model. This dataset consists of pairs of WAV files and MIDI files of piano audio, and in total, they amount to 200 hours. WAV files are raw audio files, and corresponding MIDI file contains labeled information of each WAV file, such as when, what and how strong each note was played.

We treated MIDI files as the ground truth; our model takes WAV files as the input, and trains against the MIDI files. However, both file types are hard for the model to understand, so we needed to perform data processing to begin with.

Our data processing can be broken down to two major parts. One is converting WAV files to spectrograms, and the other task is converting MIDI files to a certain kind of form which is easier to train the model.

We chose to convert MIDI files into arrays with time on the x-axis and note on the y-axis. As MIDI files hold when each note begins, how long it lasts, and how strong it was played, we divided MIDI files into three arrays respectively; first array is onsets array, the second array is frames array, and the third array is velocity array.

Since every song has a different length, we divided all the spectrograms and the labels into 20 second segments, so our model will take a 20 second long spectrogram as input.

The downside of this approach is that we might have a clip at the end of the song that is shorter than 20 seconds. This can be resolved by zero-padding, however, since most songs were several minutes long, we did not think this would make a noticeable difference and opted to simply ignore these last few seconds.

5 Architecture

The OF model[2] had seen success in the task of automatic music transcription, but since 2017 when the model was published, there have been several publications that have brought forth improvements in areas like sequence processing.

In the OF model, the processed mel-spectrogram data is encoded using a stack of convolutional layers[3]. Furthermore, the OF model uses bidirectional LSTM models as decoders. Since 2017, the transformer architecture has revolutionized the processing of sequential data[4]. Many of today's state-of-the-art models in fields like natural language processing are based on the transformer. As audio is sequential data, we think it will be advantageous to use a transformer encoder in place of the bidirectional LSTM.

Following similar data flow to OF, we will have a two-headed model with an onset and frame head. The onset prediction is passed as an input to the transformer encoder that decodes the frame predictions. We use the same loss functions for the two heads from OF (a generalization of cross-entropy loss, see eq.1-6 in[2]). For the velocity model, we use a similar approach to the frames model but without the input from onset prediction. Apart from that, we use the modified version of mean square error instead of a cross-entropy loss as a loss function. For all the models, we use Adam optimizer as our optimization function. For all the models, we also use fully connected layers in between. For a clearer picture of the model structure, please refer to figure 1 in the introduction section.

6 Baseline Model

The baseline model we compare our model to is the LSTM described in the *Onsets and Frames* paper by Google[2]. The model architecture is very simlar to figure 1, except transformers are used instead of LSTMs, and the onsets do not feed back into the LSTM.

This is a reasonable choice since our hypothesis is that introducing a transformer will improve the performance of the model. Since Google researchers likely have far greater computing resources than us and better data processing, to create a better comparison, we will write and train the LSTM model on the same processed data and machine used for the Transformer model.

7 Quantitative Results

8 Testing

One concern is that the music in the Maestro dataset share similar features. They are mostly classical pieces from the 17th to early 20th century[5], and there could be inherent biases the creators of the dataset might have when selecting the pieces and the type of recording device used.

We decided to record our own music.

9 Discussion

10 Ethical Considerations

In the United States, the music sheet industry is worth around 1 billion USD[6]. A technology that can transcribe music may render purchasing music sheets useless. Many of these sales come from music books, which often contain several scores that can be readily found online, so big publishers will likely not be greatly affected by this technology.

However, this technology will negatively affect small composers, who may make a living by selling sheet music to their music, for example on services such as MusicSpoke[6]. If this technology is accurate and readily available, it may encourage musicians to automatically generate the sheet music to the songs they like, instead of purchasing and supporting artists. Fortunately, transcribing isn't just about having the right notes, but the style in which it is presented. This is why two transcribers will not produce the same sheet music for the same song.

The MASTERO dataset and the scope of our project consist of working solely with music, so there is a low risk of discriminating against humans from pre-processing to training and post-processing. However, a large majority of the music from the dataset is Western music, so the results would likely be more reliable towards Western music. Thus, people from other cultures may not find the same success in this model as people in Western cultures do. To address this, we will attempt to find other similar datasets that were trained on music from other cultures.

References

[1] S. Hainsworth and M. Macleod, *The automated music transcription problem*, Department of Engineering, University of Cambridge, 2004.

[2] C. Hawthorne, E. Elsen, J. Song, A. Roberts, I. Simon, C. Raffel, J. H. Engel, S. Oore, and D. Eck, "Onsets and frames: Dual-objective piano transcription," *CoRR*, vol. abs/1710.11153, 2017. arXiv: 1710.11153. [Online]. Available: http://arxiv.org/abs/1710.11153.

[3] R. Kelz, M. Dorfer, F. Korzeniowski, S. Böck, A. Arzt, and G. Widmer, "On the potential of simple framewise approaches to piano transcription," *CoRR*, vol. abs/1612.05153, 2016. arXiv: 1612.05153. [Online]. Available: http://arxiv.org/abs/1612.05153.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *CoRR*, vol. abs/1706.03762, 2017. arXiv: 1706.03762. [Online]. Available: http://arxiv.org/abs/1706.03762.

[5] C. Hawthorne, A. Stasyuk, A. Roberts, I. Simon, C.-Z. A. Huang, S. Dieleman, E. Elsen, J. H. Engel, and D. Eck, "Enabling factorized piano music modeling and generation with the MAESTRO dataset," *CoRR*, vol. abs/1810.12247, 2018. arXiv: 1810.12247. [Online]. Available: http://arxiv.org/abs/1810.12247.

[6] M. Hunckler, *Musicspoke looks to disrupt 1 billion sheet music industry with marketplace for artist-owned scores*, Forbes, 2017.