MUSIC TRANSCRIPTION USING IMAGE SEGMENTATION

Zach Chun, Tommy Merth, Althea Poteet, Vivaan Bhatia

Introduction

Music can be categorized as **monophonic** (single note) and **polyphonic** (multi-note). Monophonic music transcription is a solved problem[1] whereas polyphonic transcription presents some fundamental challenges, such as overlapping frequencies. Since most music is polyphonic, the ability to automatically transcribe it is important for three reasons:

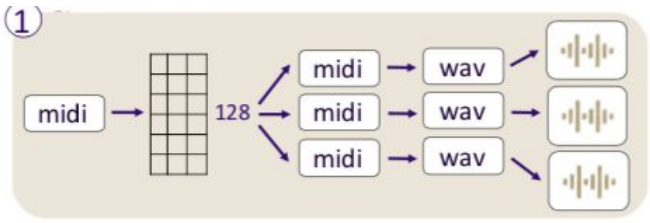
1. Most people are not ear trained, and find it difficult to learn a song just by listening to it. Especially with polyphonic music! Midi files and sheet music simplifies learning to play songs
2. A transcribed version of a song makes it easier to analyze and modify
3. MIDI (Musical Instrument Digital Interface) files are useful for Digital Audio Workstations (DAWs), like GarageBand

Investigation and Inspiration

Most pre-existing models we encountered during our research converted MIDI files into spectrograms and performed multiclass classification per quarter note, eighth note, etc. We thought this approach was inflexible since it requires the music to be downsampled to some predetermined interval.

Instead of taking this approach, we wanted to directly infer the notes with as much precision as necessary. Given a spectrogram, we can think of this transcription task in the exact same way as image segmentation.

Data Generation

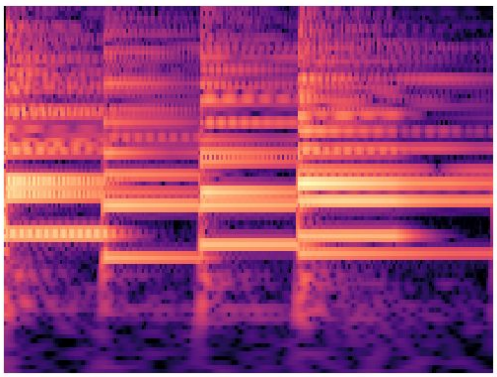


So far we have only generated datasets of classical music. We obtained ~40 full midi files from an online MIDI database[2] and synthetically generated .wav files using an audio plugin called Fluidsynth. From each ~5 minute long .wav file, we generate 5 second spectrograms from a single MIDI file using the constant Q transform.

[1] https://core.ac.uk/download/pdf/48626932.pdf
[2] http://www.piano-midi.de/midi_files.htm

Some of this preprocessing code was borrowed from another student group[3].

Model Details



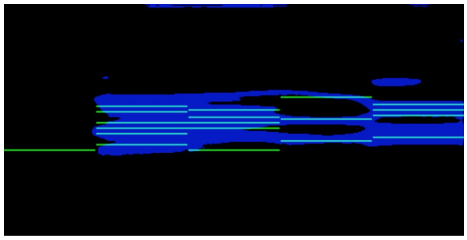
Since we have processed our data in such a way that the inputs and labels have the same dimensionality, we can easily reduce this problem to image segmentation. We used UNet[4] as our initial model, but reduced its complexity by removing two encoder and decoder layers. Since we are predicting a one-hot encoding, we would eventually need to implement smoothing/post-processing to extract the final midi encoding.

Since our label distribution is very unbalanced (~45:1), we used the weighted binary cross entropy loss as our loss function, penalizing false negatives by a factor of ~20-40.

Results

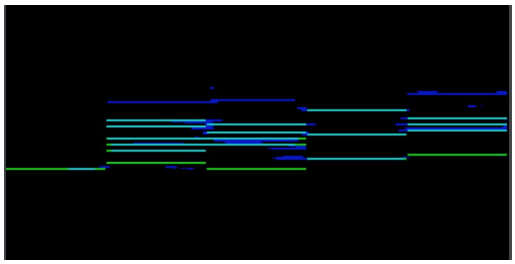
The extremely high accuracy of our model (> 99%) is not meaningful. Instead, we used the F1 score to evaluate our model. We achieved over 0.7 for this metric without extensive hyperparameter tuning.

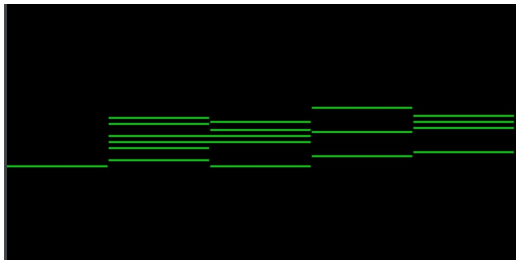
Interestingly, the model occasionally made an error in transcription by predicting the octave of a correct note instead of that note itself. We believe this is evidence that the model learned the relationship between a note and the frequencies it generates (as opposed to simply performing some sort of max-pooling on the spectrogram).



[Prediction at Epoch 3]

[3] https://github.com/chaumifan/DSL_Final
[4] https://github.com/EKami/carvana-challenge/blob/master/src/nn/unet.py


[Prediction at Epoch 10]


[Midi To Replicate]

Above we have included three images that showcases our results. The last image, is the MIDI we tried to replicate in our test set. The first and second images are our predictions, where the blue areas are where we predicted notes, green is what we missed, and the overlap is what we predicted correctly.

We can see that around epoch ten, our model started to learn the height distribution for the placement of these pixels as well as the precise location. Since we took more of a computer vision approach when solving this problem, we are happy with the results.

Conclusion

This proof-of-concept shows that image segmentation is a valid approach to polyphonic music transcription. With a more comprehensive dataset, hyperparameter tuning, and different initializations for our weights, we believe that one can achieve an F1 score that is competitive with state-of-the-art approaches.

We aren't claiming that this is necessarily the best approach. However, since we have reduced transcription to a very pertinent problem in computer vision, this approach will benefit from any advances in that field.