Neural Synthesis of Piano Performance

Patricia Lan, Steve Leung, Grant Yang

Introduction

Performing music from musical scores is an ill-conditioned problem. The audio representation of a performance contains significantly more information than the original score. Musicians must make decisions about tone, tempo, and dynamics based upon complex non-causal relationships between the musical notes, phrases, and sections. Further, the same piece may have many equally valid and pleasing interpretations.

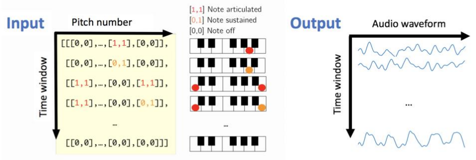
In this study, we trained a bidirectional RNN to “play" music by learning the relationship between musical scores and recordings of virtuoso musicians.

Data

We used the Piano Dataset curated by Malik et al. This dataset is comprised of 349 classical piano performances encoded in MIDI file format. We used the MIDI files and GarageBand to create audio files of the performances. The data was then reformatted for the RNN.

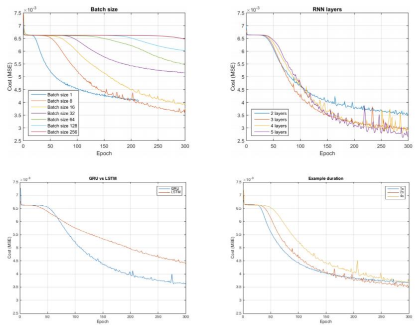
Input:
Pitch encoding matrix [examples, time windows, 176]

Output:
Audio matrix [examples, time windows, audio samples]



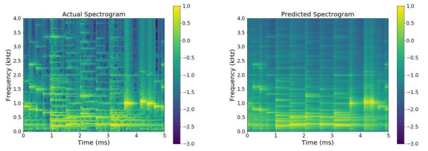
Bidirectional RNN

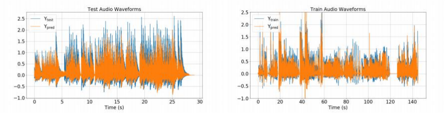
We tested multiple hyperparameters, architectures, and input data formats to improve model training. The options included batch size, number of RNN layers, RNN cell type, and example duration.

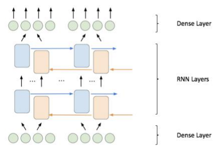


Results

The audio files can be found at
https://tinyurl.com/y8zo3mqz







Discussion

Using the log magnitude spectrogram loss caused the model to train drastically faster than when using the time domain MSE loss. In fact, using the new cost function made the largest difference - it was more influential than tuning the hyperparameters and network architecture.

The sequential nature of RNN training made training slow to compute and hard to parallelize. GPUs did not speed up training because a lot of time was spent on I/O operations (due to our batch size of 8).

Future directions

The loss of phase information was a major problem with achieving natural sounding audio. In the future, we would like to explore ways to include phase information in our loss function. We would also include the use of the sostenuto pedal into the input of our next model. The sostenuto pedal allows notes to continue after the keys are released and can have a dramatic effect on the timbre and temporal evolution of the piano sound. Finally, we would like to train our model on different instruments.

References

Malik et al. https://arxiv.org/abs/1708.03535
Graves. https://arxiv.org/abs/1308.0850
Fu et al. https://arxiv.org/pdf/1704.08504.pdf