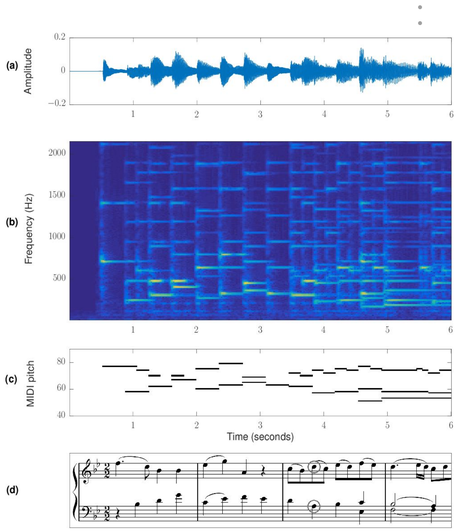
Piano Transcription with Transformers

Jerry Cheung, Andrew Li, Matthew Zhou

Georgia Institute of Technology

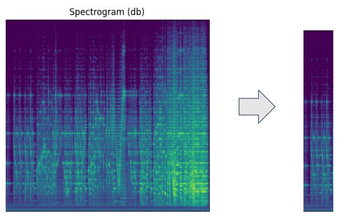
With the rapid development of deep neural networks, much progress has been made in Automated Music Transcription (AMT), the converting of audio waveforms to musical notation. We apply the Transformer model, which has been shown to be an improvement upon RNNs generally, and we propose a novel model that predicts the presence of note onsets and offsets in conjunction with direct transcription, which we find to be an improvement on a plain transformer baseline.

Problem Definition

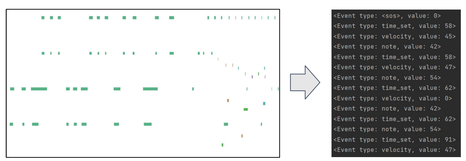


Dataset

We use the MAESTRO V3.0.0 dataset, consisting of 1276 WAV and MIDI files of virtuosic piano performances of classical music. We transform the WAV files into log mel spectrograms which we then split into segments 4.088 seconds long.



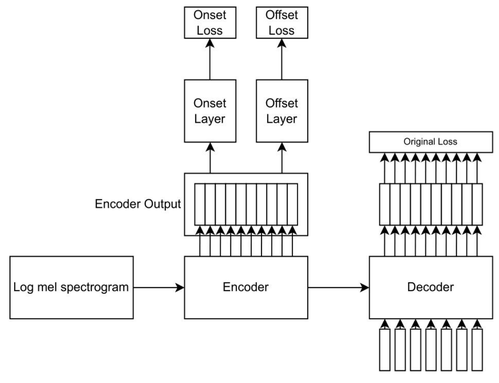
We process the MIDI files into events of note onset and offset times, note pitches, and note velocities. We split these as well to match the audio segments



These spectrogram and event segments are then fed to our model's encoder and decoder respectively for training.

Method

- We adapt the traditional transformer architecture, with 8 attention heads, 5 encoder layers and decoder layers and a feed-forward layer of dimension 1024.
- We introduce an onset layer and offset layer to the transformer architecture for an aggregated loss function.
- Through making the model predict onsets and offsets from the encoder, we hope this objective encourages a more informative hidden state from the encoder to the decoder for this transcription task.



Experiments & Results

Model	Onset F1	Onset & Offset F1	Onset, Offset, & Velocity F1
Transformer Baseline	69.43	62.20	58.18
Proposed Method	71.20	64.32	58.29

Our proposed model has a much higher Onset F1 score and Onset & Offset F1 score. The addition of the onset and offset layers allows the model's encoding layer to more accurately represent the onsets and offsets of the notes. This also means that the Onset, Offset, & Velocity F1 is only slightly better for our proposed model, since there is less information captured by the encoder regarding the velocity of the notes.

Conclusions

We have shown that a generic Transformer with limited computation resources can achieve good performance mapping spectrograms to MIDI-like output events.

With small adaptations to the transformer model, we are able to further increase its performance. Our experiments show that there are still room for improvement for future works. We wish to explore the following ideas in the future:

- Use of sparse attention to enable transcription of an entire piece of music in a single pass.
- Distillation of similar models on more resource constrained devices such as mobile phones.