Homework #3: Automatic Polyphonic Piano Transcription

20204310 Haemin Kim

Experiments & Results

Polyphonic piano transcription aims to automatically convert piano musical signals into a piano roll. In addition to the baseline model implemented in CNN, RNN-based networks are implemented and tested. Please refer to **model.py** for the actual code. All results are trained with default options except for experiment 4.

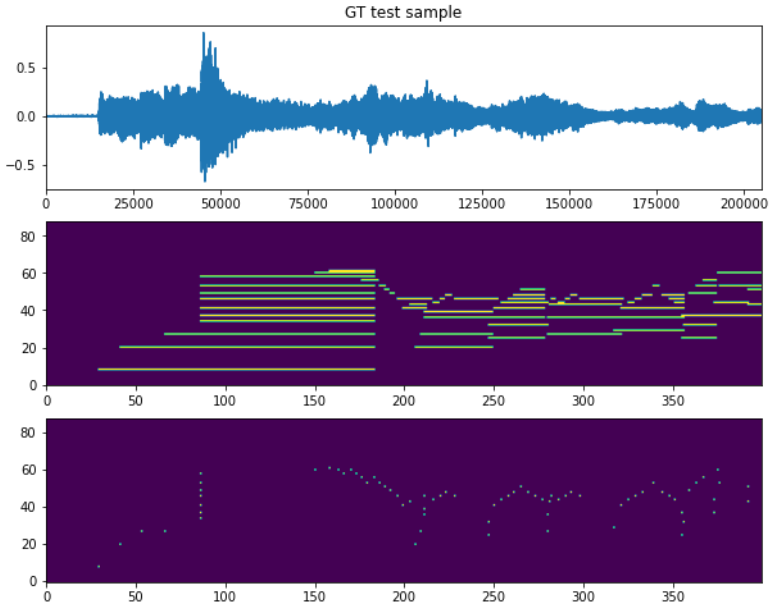
Options

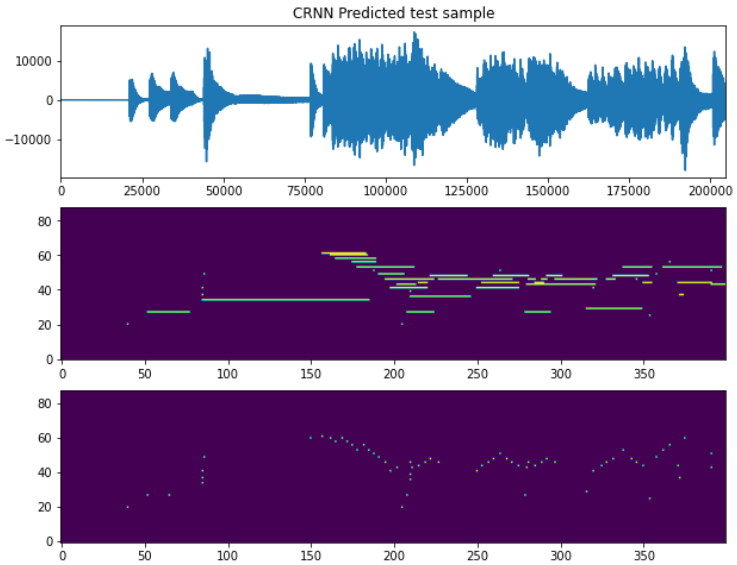
- sequence_length : 102400
- learning_rate : 6e-4
- batch_size : 16
- iterations : 10000
- validation_interval : 1000
- weight_decay : 0
- cnn_unit : 48
- fc_unit : 256
- Constants in constants.py are left unchanged

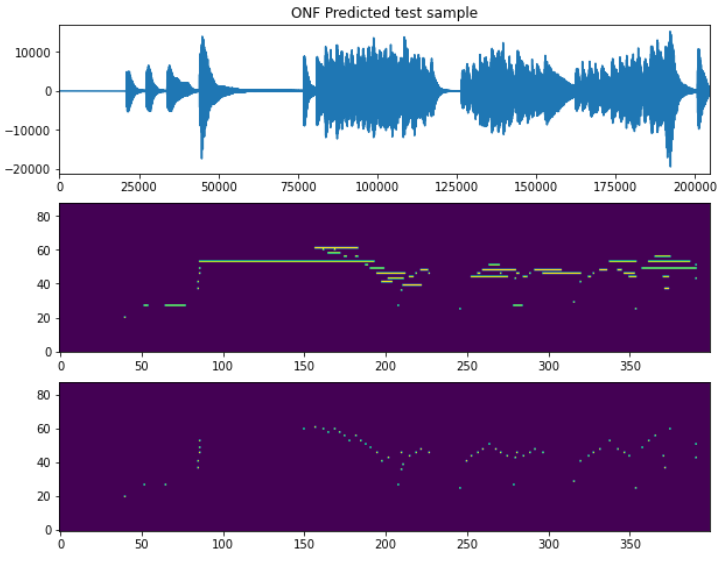
Neural Network for Piano Transcription						
#	Model	frame loss / onset loss	frame F1	onset F1	note F1	note w/ offset F1
1	RNN	0.122 / 0.012	0.503	0.431	0.502	0.164
2	CRNN	0.118 / 0.008	**0.547**	0.692	0.794	**0.326**
3	ONF	0.126 / 0.008	0.439	**0.704**	**0.800**	0.294
4	ONF w/ win_length = 1024	0.129 / 0.009	0.430	0.654	0.744	0.270

The model architectures that consist of both CNN and LSTM showed better results than the one with only LSTM layers. When CRNN and ONF models are compared, the CRNN predicted frames better than ONF and ONF was better regarding note-level predictions. It can be inferred that using the encoded onset information through causal connection improved note prediction.

[MIDI-Unprocessed_03_R1_2006_01-05_ORIG_MID--AUDIO_03_R1_2006_04_Track04_wav]



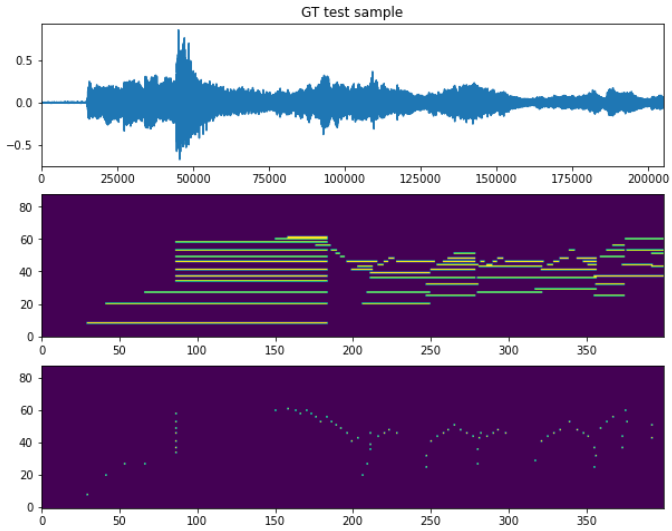


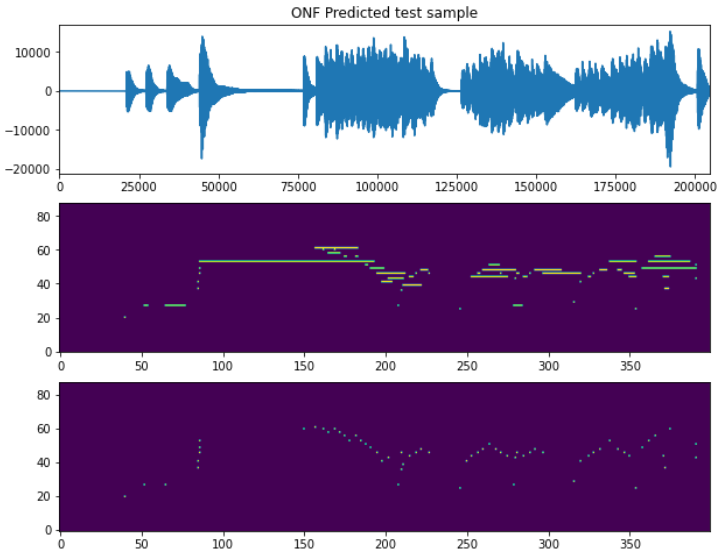


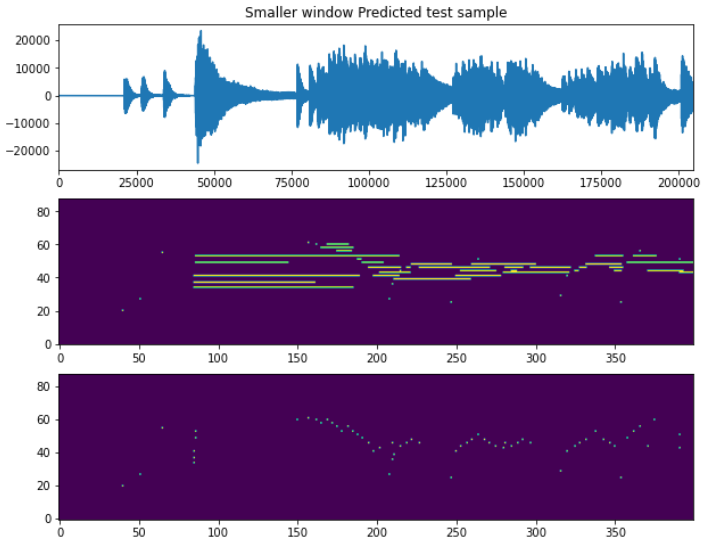
Experiment 4 was conducted to check if the time resolution and frequency resolution of the input Mel Spectogram affects the performance of the ONF model. The original window length of ONF was 2048 and it was changed to 1024 in the experiment. By decreasing the window size, the time resolution is increased and the model is expected to predict timing information better while losing some performance in predicting frequency or notes.

Given the results, the original ONF with window size of 2048 still showed better f1 scores than ONF with window size of 1024. Although increased time resolution did not affect the results from the metrics, a test sample visualization shows that it predicted sustain of a note better than the former one. Therefore, when the two samples are played, the one predicted with a smaller window is considered to be more natural music played by humans.

[MIDI-Unprocessed_03_R1_2006_01-05_ORIG_MID--AUDIO_03_R1_2006_04_Track04_wav]



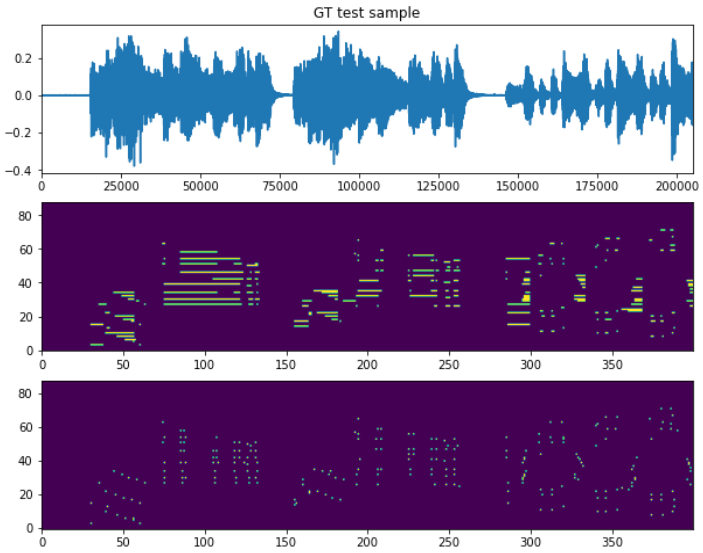


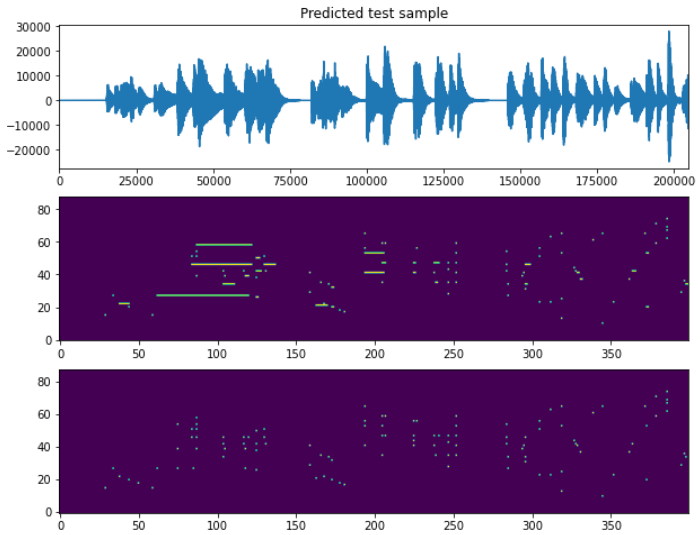


Discussions

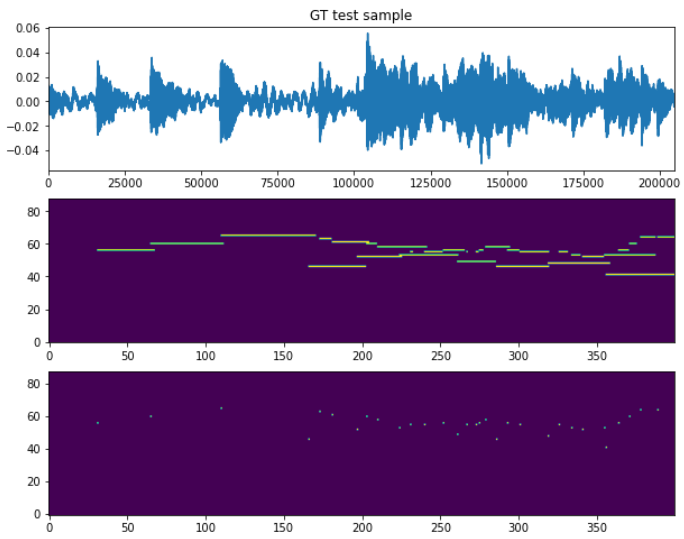
Two outputs from ONF are visualized along with the predicted piano rolls. As seen in the visualizations, many predicted onsets and notes match with that of ground truth data, producing similar audio signals in midi format. Also, onset predictions mostly match with the frame prediction made by the same model.

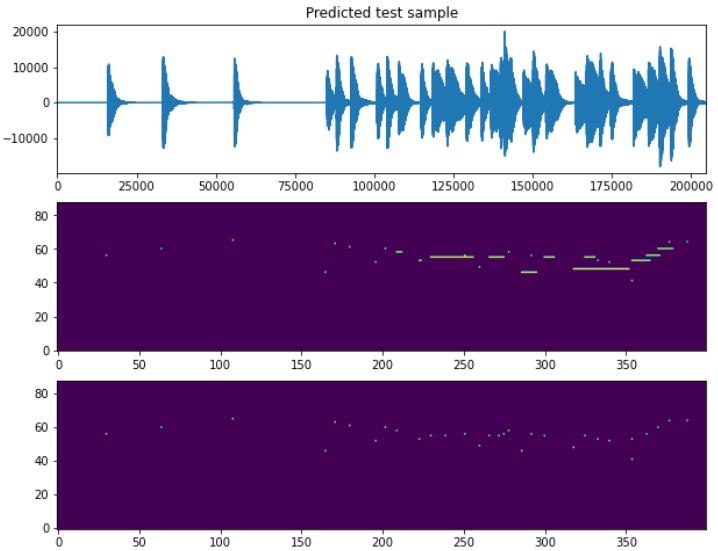
[ORIG-MIDI_01_7_7_13_Group__MID--AUDIO_11_R1_2013_wav--4.flac]





[MIDI-Unprocessed_Recital9-11_MID--AUDIO_09_R1_2018_wav--5.flac]





However, there are still a few errors in the prediction. The model struggles to capture long term sustain in frame-level predictions and rather falsely predict that there are additional onsets at that time. Also, the part where multiple notes have to be played simultaneously is not predicted precisely. The notes are slightly misaligned and give different impressions when played compared to the ground truth music.

To improve the performance of the automatic polyphonic piano transcription, it is important to refine the onset prediction by playing correct notes at the correct time. Therefore, some encoded beat information could be considered. Just like how onset information is used for frame prediction in ONF, beat tracking information extracted from language models could be used to predict exact timing of each onset. On top of beat tracking, chord recognition could also be adopted to aid the model to recognize every note played at the same beat. Lastly, new network architecture could be explored such as transformers. Considering the fact that automatic music transcription tasks benefited from sequential models found in the field of natural language processing, transformers would be a good choice to speed up the training time and achieve superior results.