Homework 3: Automatic Polyphonic Piano Transcription

20213013 Jiho Kang

November 22th, 2021

[Experiments and Results]

The experiment was conducted under the same conditions (given settings) for all models (baseline, rnn, crnn, onf) except for iteration. The results are shown in Tables 1,2.

		Baseline	RNN	CRNN	ONF
Frame	**Frame Precision**	0.814 ± 0.049	0.781 ± 0.081	0.681 ± 0.059	0.632 ± 0.062
	Frame Recall	0.559 ± 0.094	0.232 ± 0.061	0.483 ± 0.081	0.394 ± 0.090
	Frame F1	0.658 ± 0.067	0.354 ± 0.074	0.562 ± 0.067	0.482 ± 0.079
	Onset Precision	0.833 ± 0.025	0.724 ± 0.042	0.814 ± 0.035	0.813 ± 0.036
	Onset Recall	0.660 ± 0.117	0.314 ± 0.140	0.601 ± 0.127	0.597 ± 0.128
	Onset F1	0.732 ± 0.080	0.422 ± 0.132	0.686 ± 0.094	0.683 ± 0.095
Note	**Precision**	0.986 ± 0.007	0.928 ± 0.030	0.972 ± 0.014	0.967 ± 0.015
	Recall	0.720 ± 0.117	0.348 ± 0.147	0.659 ± 0.129	0.655 ± 0.130
	F1	0.827 ± 0.079	0.490 ± 0.149	0.779 ± 0.096	0.775 ± 0.097
	Overlap	0.575 ± 0.054	0.370 ± 0.085	0.492 ± 0.063	0.437 ± 0.086
Note with Offset	**Precision**	0.492 ± 0.115	0.275 ± 0.148	0.428 ± 0.114	0.379 ± 0.135
	Recall	0.363 ± 0.119	0.107 ± 0.082	0.291 ± 0.102	0.256 ± 0.110
	F1	0.415 ± 0.117	0.149 ± 0.104	0.344 ± 0.106	0.303 ± 0.119
	Overlap	0.863 ± 0.058	0.787 ± 0.083	0.850 ± 0.074	0.836 ± 0.074

Table 1. Precision, Recall, and F1 results on test dataset (iteration value is 10,000)

		Baseline	RNN	CRNN	ONF
Frame	**Frame Precision**	0.785 ± 0.051	0.740 ± 0.064	0.717 ± 0.049	0.657 ± 0.059
	Frame Recall	0.609 ± 0.097	0.423 ± 0.088	0.636 ± 0.076	0.499 ± 0.079
	Frame F1	0.680 ± 0.062	0.533 ± 0.079	0.672 ± 0.054	0.565 ± 0.065
	Onset Precision	0.832 ± 0.027	0.761 ± 0.039	0.823 ± 0.035	0.816 ± 0.035
	Onset Recall	0.679 ± 0.114	0.395 ± 0.139	0.686 ± 0.110	0.681 ± 0.114
	Onset F1	0.744 ± 0.076	0.508 ± 0.121	0.745 ± 0.078	0.739 ± 0.080
Note	**Precision**	0.988 ± 0.006	0.947 ± 0.024	0.974 ± 0.013	0.970 ± 0.015
	Recall	0.741 ± 0.113	0.440 ± 0.147	0.757 ± 0.111	0.751 ± 0.114
	F1	0.842 ± 0.075	0.587 ± 0.134	0.848 ± 0.075	0.842 ± 0.078
	Overlap	0.597 ± 0.052	0.459 ± 0.050	0.588 ± 0.047	0.495 ± 0.062
Note with Offset	**Precision**	0.514 ± 0.115	0.327 ± 0.112	0.525 ± 0.091	0.432 ± 0.114
	Recall	0.389 ± 0.119	0.157 ± 0.088	0.409 ± 0.099	0.335 ± 0.109
	F1	0.440 ± 0.117	0.208 ± 0.099	0.458 ± 0.094	0.376 ± 0.110
	Overlap	0.870 ± 0.058	0.823 ± 0.079	0.871 ± 0.068	0.852 ± 0.072

Table 2. Precision, Recall, and F1 results on test dataset (iteration value is 20,000)

[Discussion]

1. Visualize at least one sample of your prediction (onset and frame) in the piano roll format

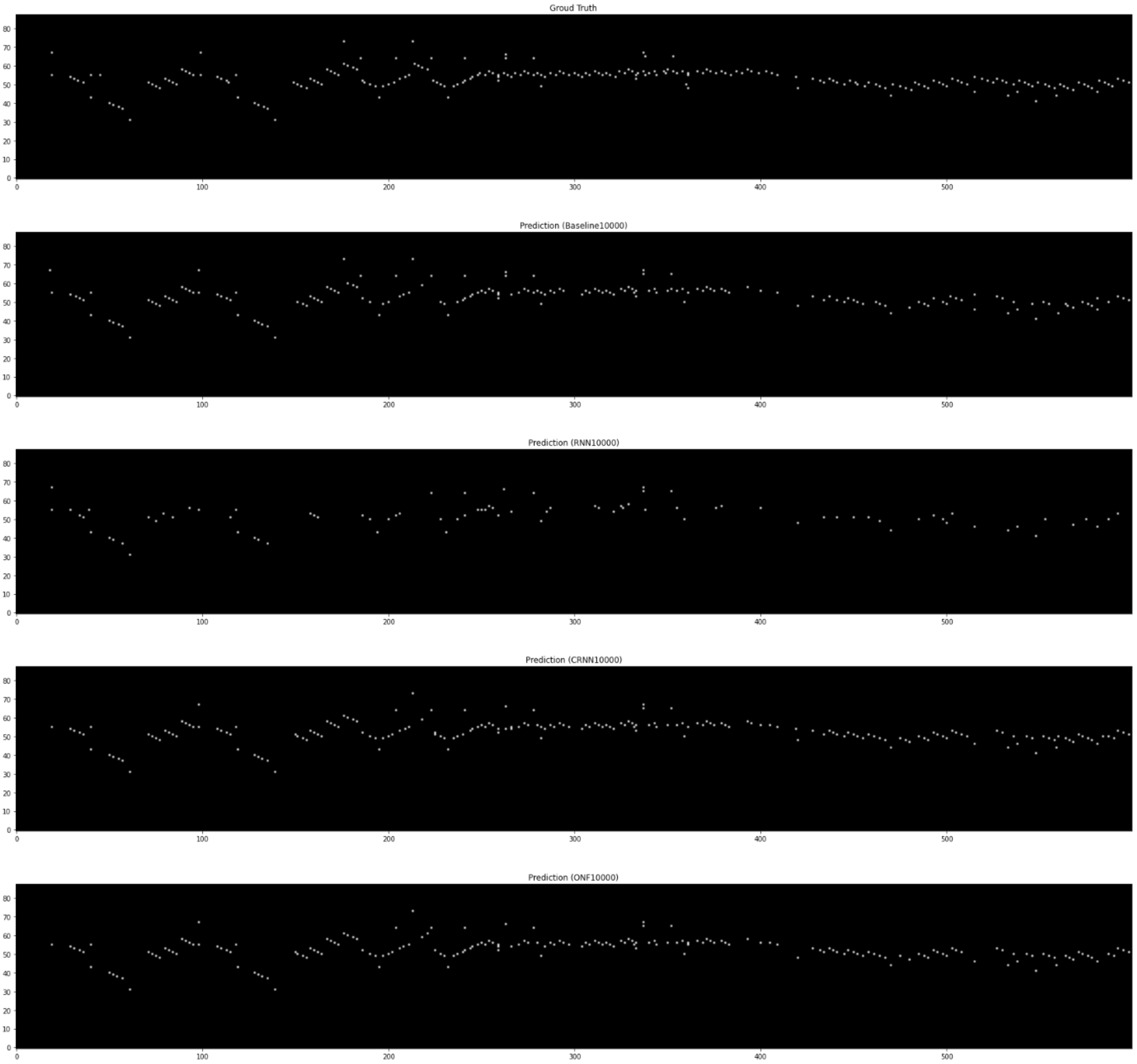
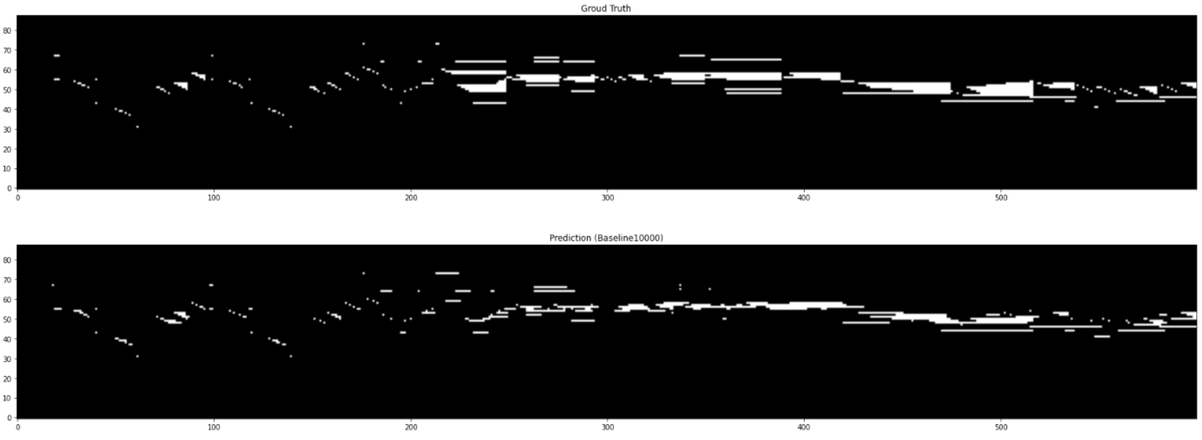


Figure 1. Visualization of onset (ground truth and prediction). A Part of the file 'MIDI-Unprocessed_SMF_17_R1_2004_03-06_ORIG_MID--AUDIO_20_R2_2004_12_Track12_wav' file'



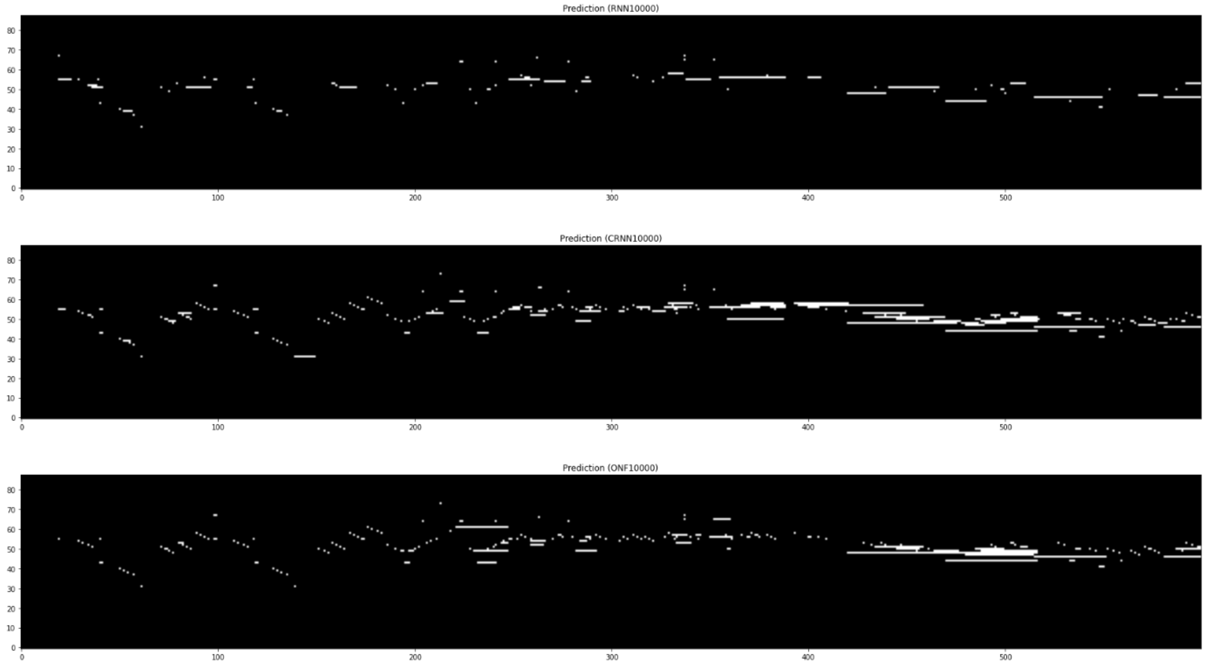


Figure 2. Visualization of frame (ground truth and prediction). A Part of the file 'MIDI-Unprocessed_SMF_17_R1_2004_03-06_ORIG_MID--AUDIO_20_R2_2004_12_Track12_wav'

2. What kinds of errors did you observe?

A. Are the predicted onsets and frames consistent with each other?

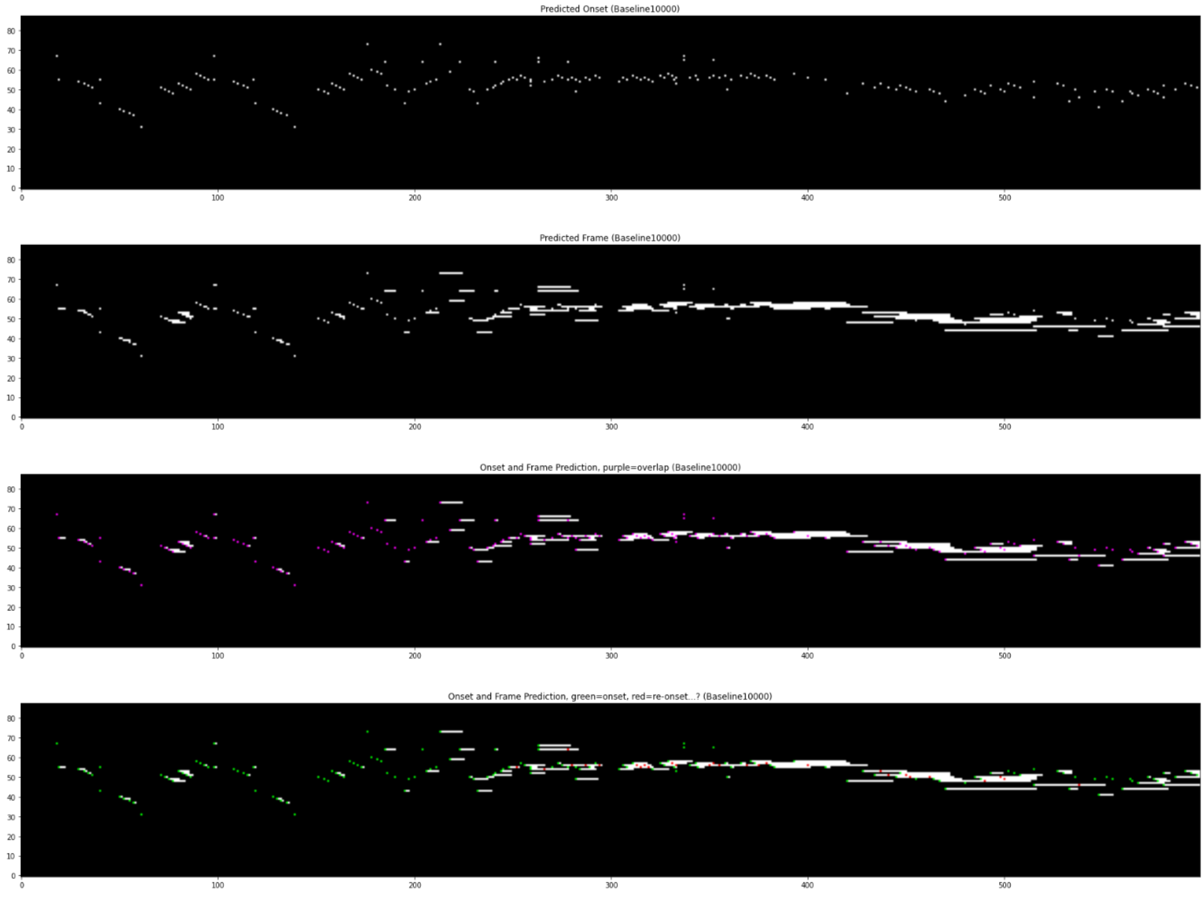


Figure 3. Visualization of predicted onset and frame respectively, the result of overlapping onset and frame (Baseline)

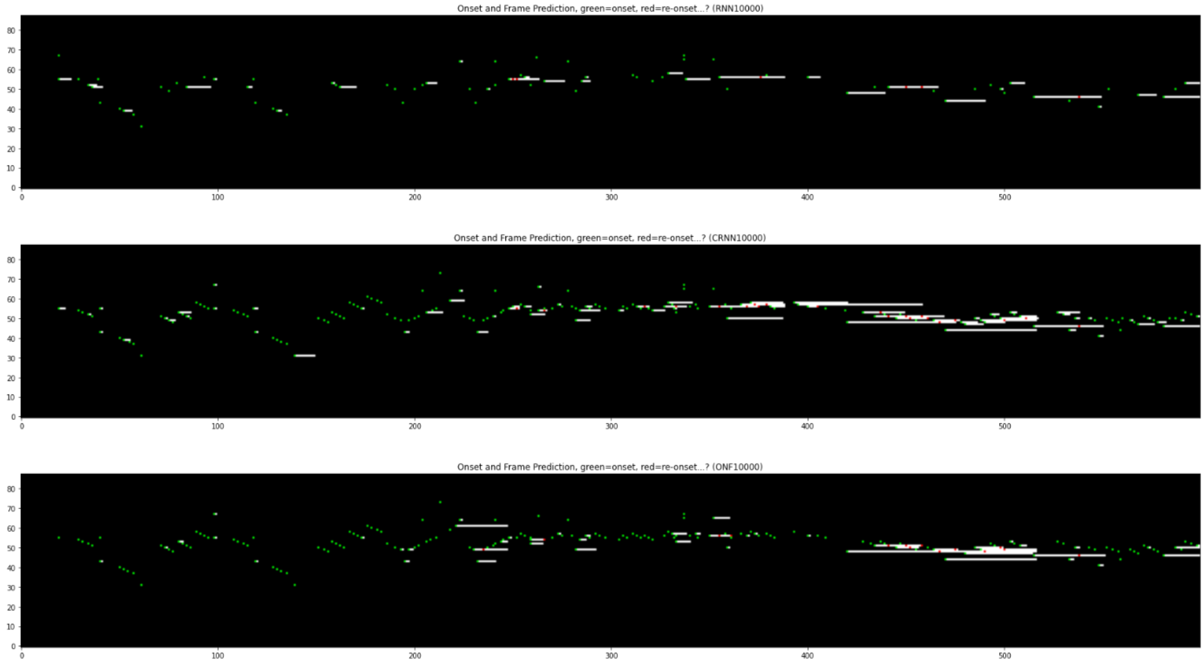
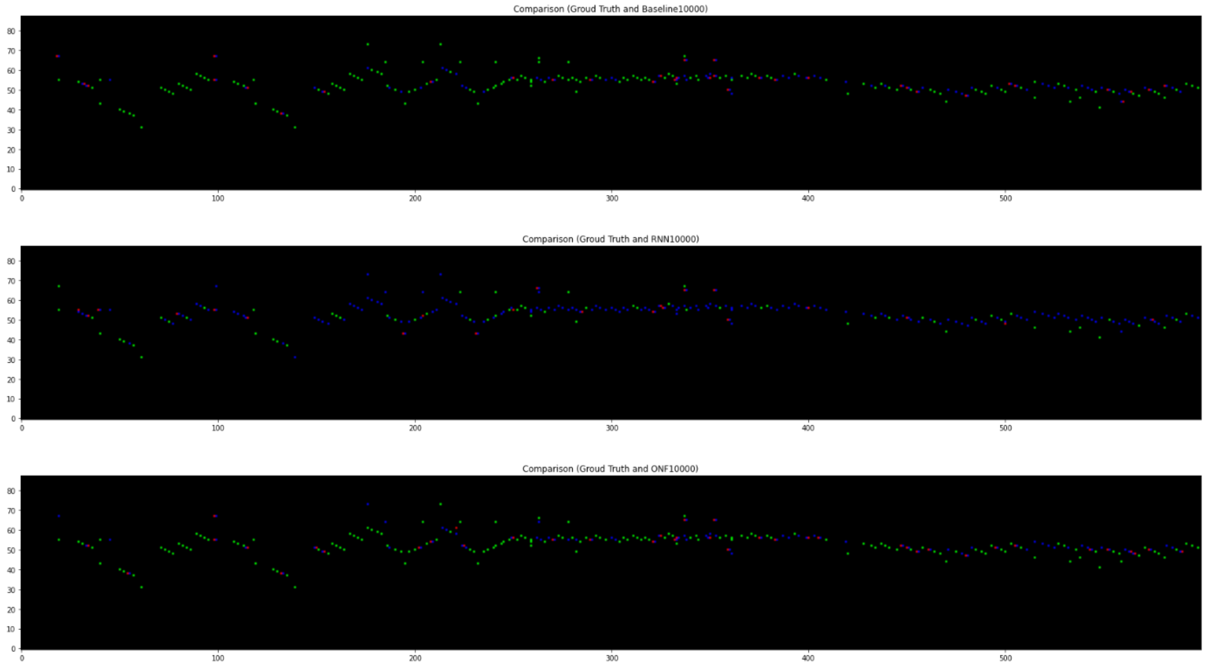


Figure 4. The result of overlapping onset and frame (RNN, CRNN, ONF)

As can be seen from Figures 3 and 4, all models showed that predicted onsets and frames consistent with each somewhat. One question is that the onsets appear again before the frame is turned off in addition to moment when the frame is turned on, as shown in figure 9 where the onset of the ground truth overlaps with the frame of the ground truth is displayed in purple. To distinguish these, the onset at the moment the frame is turned on is displayed in green, and the other onsets (can be interpreted as a re-onset) are displayed in red (figure 3,4,9).

B. Compare them with the ground truth and analyze the errors in both frame-wise and note-wise perspective.



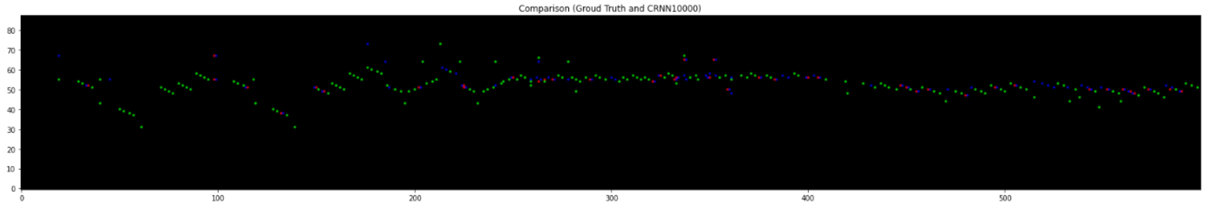


Figure 5. Visualization of onset (Comparison between ground truth and prediction). True Positive (Green), False Positive (Red), False Negative (Blue)

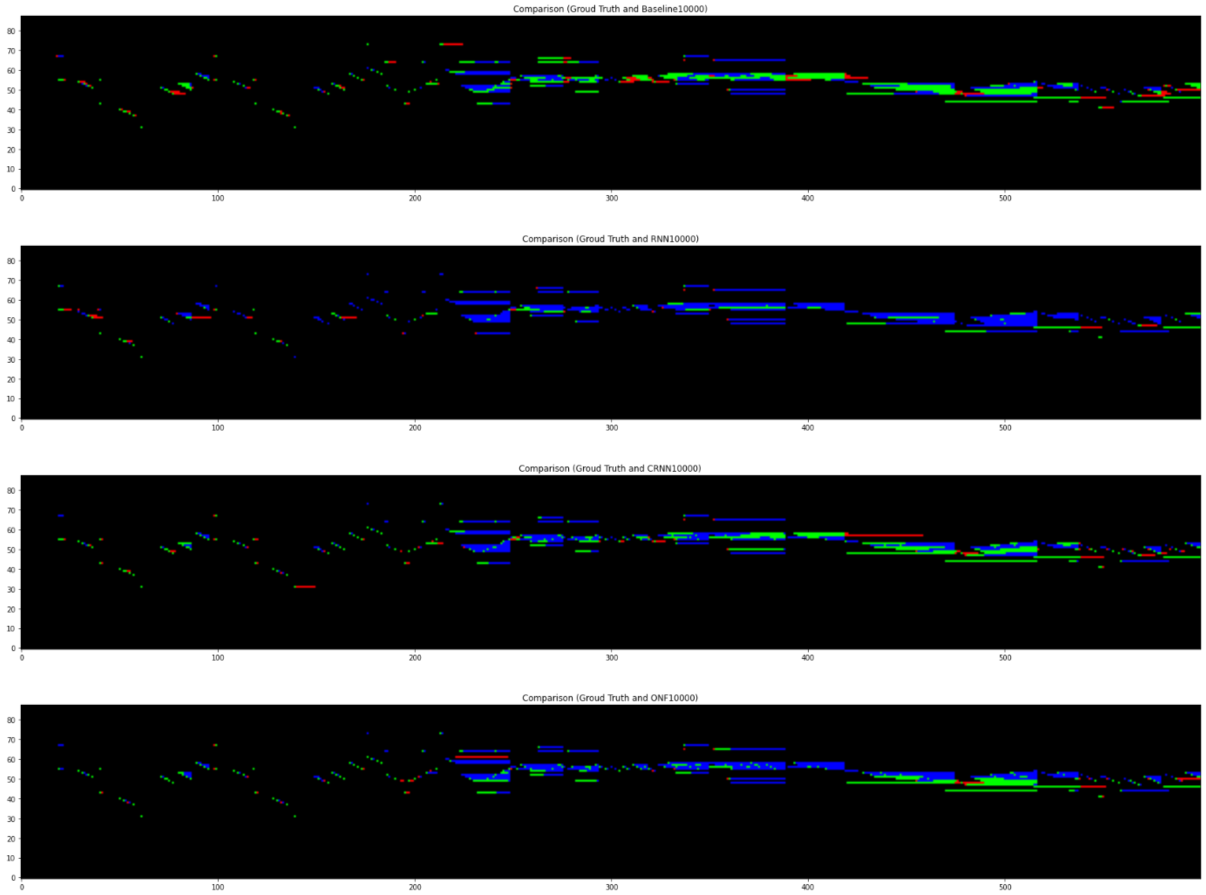


Figure 6. Visualization of frame (Comparison between ground truth and prediction). True Positive (Green), False Positive (Red), False Negative (Blue)

To compare predicted onsets and frames with the ground truth and analyze the errors in both frame-wise and note-wise perspective, the visualization was conducted in the way as figures 5 and 6. In figures, green pixels represents true positive, blue pixels represents false positive, and red pixels represents false. For both aspects' onset and frames, contrary to expectations, Baseline performed the best, CRNN and ONF performed similarly, and RNN performed the worst as shown in figure 5,6 and table 1 (iteration value is set to 10,000). One notable point is that the proportion of false negatives in frame prediction is high. In other words, models cannot predict multiple musical notes for each column on the horizontal axis. Therefore, it can be considered that the proposed model structures of this assignment do not capture the synchronicity of musical notes properly.

3. How would you improve the results?

The key point in this assignment is separation of onset state detection part in network. A connection from the onset prediction to the input of frame prediction network restrict prediction framewise detector [1]. This is what empowers note state in a polyphonic piano music transcription task. According to [3], note-level accuracy is more important than the frame-level accuracy in terms of human musical perception. In line with this aspect, the following direction can be investigation about more note states such as onset, sustain, offset, re-onset, and even detection of the sustain pedal. To this end, we can make use of unified neural network architecture which can predict multiple note states using a soft-max output with a single loss function [2].

Reference

[1] Curtis Hawthorne, Erich Elsen, Jialin Song, Adam Roberts, Ian Simon, Colin Raffel, Jesse Engel, Sageev Oore, and Douglas Eck. Onsets and frames: Dual-objective piano transcription. In Proceedings of the 19th International Society for Music Information Retrieval Conference, 2018.

[2] T. Kwon, D. Jeong, and J. Nam, “Polyphonic piano transcription using autoregressive multi-state note model,” in Proceedings of the 21st International Society for Music Information Retrieval Conference (ISMIR), 2020.

[3] GCT634/AI613: Musical Applications of Machine Learning (Fall 2021) ppt slide, “Automatic Music Transcription”

Appendix

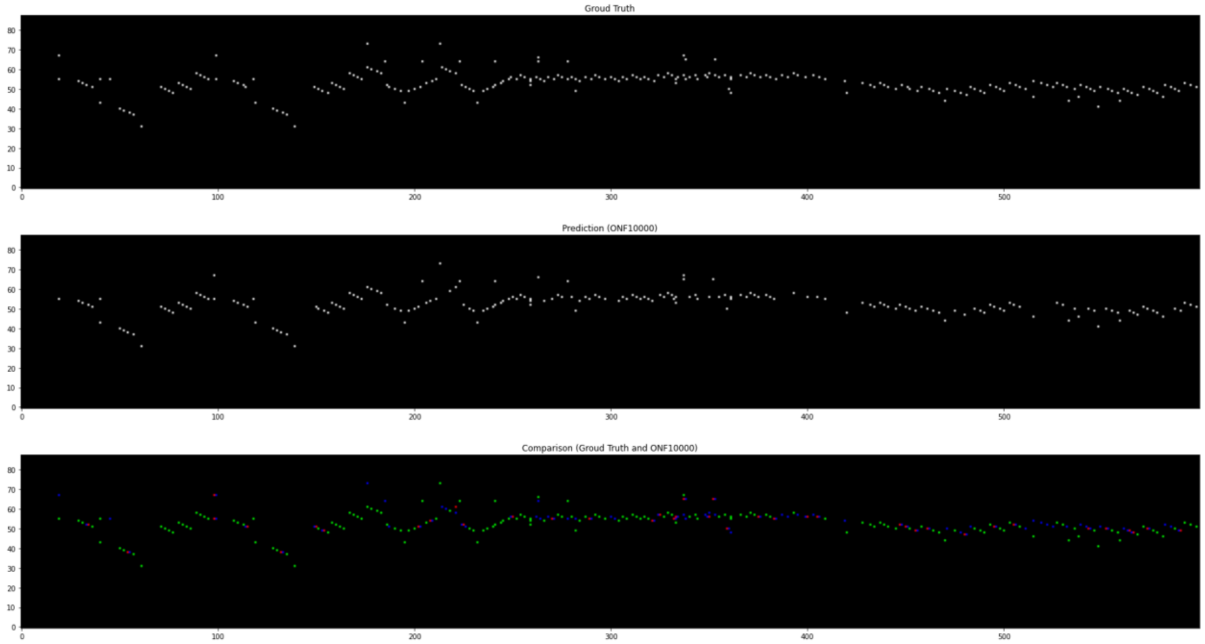


Figure 7. Visualization of onset (ground truth, prediction, and comparison)

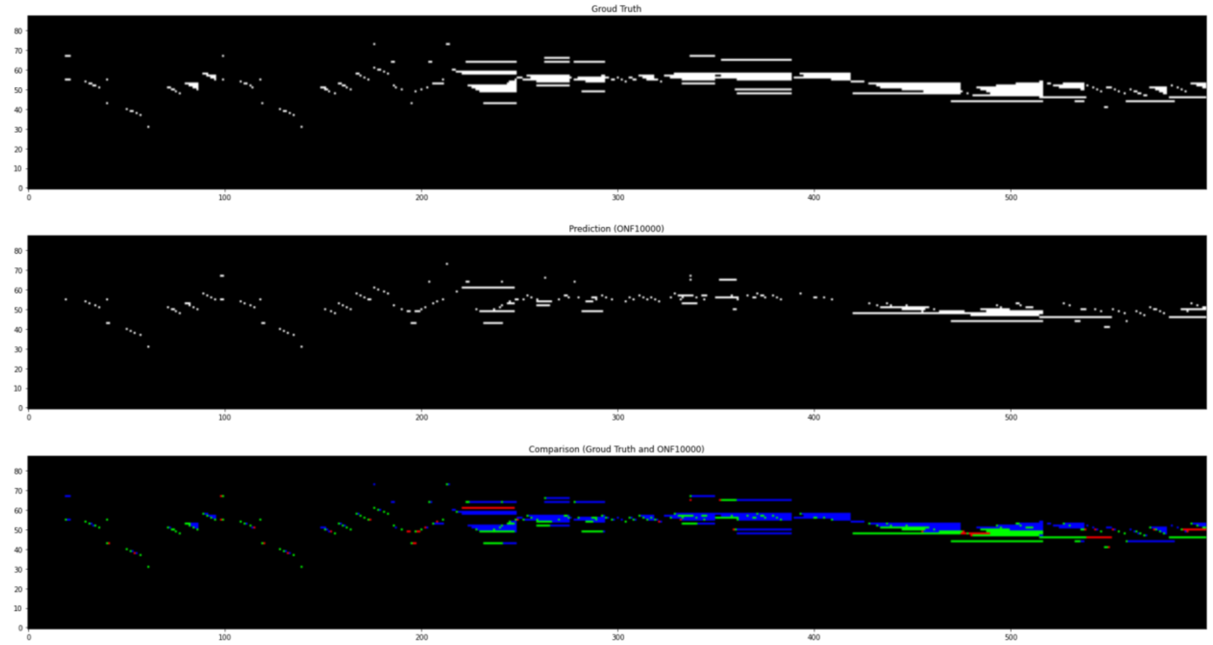


Figure 8. Visualization of frame (ground truth, prediction, and comparison)

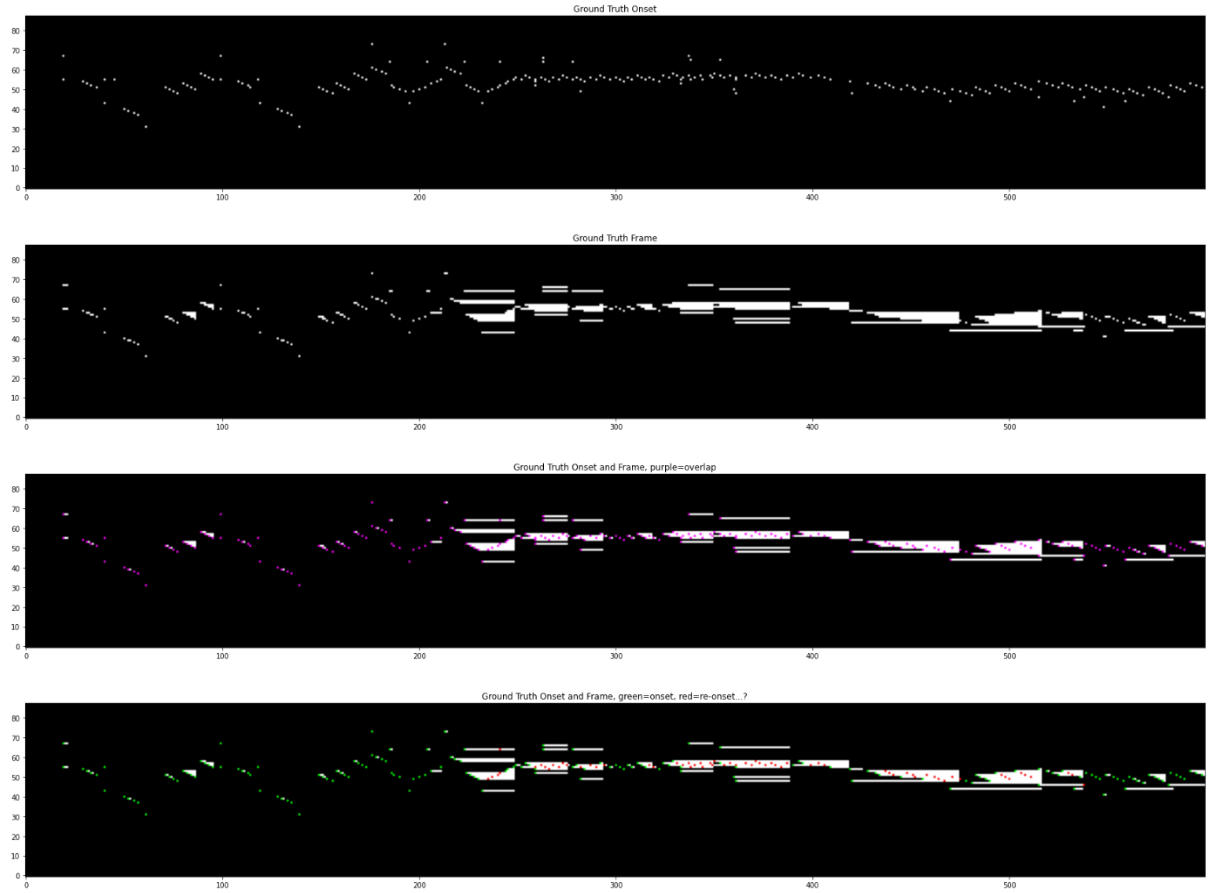


Figure 9. Visualization of ground truth onset and frame respectively, the result of overlapping onset and frame