Experiment Report: Joint Multi-pitch Detection and Score Transcription for Polyphonic Piano Music

Lele Liu, Veronica Morfi, and Emmanouil Benetos

1 Introduction

This report provides additional experimental details for paper:

- Lele Liu, Veronica Morfi, and Emmanouil Benetos, "Joint Multi-pitch Detection and Score Transcription for Polyphonic Piano Music," submitted to ICASSP 2021.

Section 2 includes detailed results on comparing different input representations and their parameters. Results on the time and memory usage for different score representations are described in Section 3.

2 Comparison of time-frequency representations

In this experiment, we compare different input representations in the form of audio spectrograms for multi-pitch detection, including Short Time Fourier Transform (STFT), Mel Spectrogram, Constant-Q Transform (CQT), Harmonic Constant-Q Transform (HCQT) and Variable-Q Transform (VQT). Please refer to the original paper for parameter and model choices. Model performance is evaluated using the benchmark frame-level and note-level evaluation metrics for automatic music transcription in MIREX [1].

Here, we provide additional results on the performance with different parameters and the precision/recall scores for the metric in Table 1. Results show what the overall best input representation is the VQT spectrogram with 60×8 frequency bins and a gamma value of 20. We discover some trend in how the parameters influence model performance. For example, a window length of 2048 outperforms 1024 for STFT and Mel Spectrogram. Larger number of frequency bins tend to result in higher performance for Mel Spectrogram, CQT, HCQT and VQT.

Figure 1 includes example ground truth and transcribed piano rolls using the best input representation. Generally, the pitches are correctly predicted. However, the model is not very good at predicting short gaps between repeated notes, which results in merged note errors.

3 Comparison of score representations

We compare two score representations for symbolic score output: 1) LilyPond representation and 2) Reshaped representation. Details about those two score representations and their performance in terms of Word Error Rate [2] and MV2H metric [3] are included in the original paper. Here, we provide in Table 2 additional information on the time and space resources required by the two score representations. Training time/memory is measured as average values per epoch with a batch

Table 1: F-measure of piano-roll prediction on different input representations and parameters. N_w: window length, N_m: number of mel bands, N_b: number of bins per octave, N_o: number of octaves, N_h: number of harmonics. P_f: framewise precision, R_f: framewise recall, F_f: framewise F-score, P_{on}: notewise onset only precision, R_{on}: notewise onset only recall, F_{on}: notewise onset only F-score, P_{onoff}: notewise onset and offset precision, R_{onoff}: notewise onset and offset recall, F_{onoff}: notewise onset and offset F-score.

Input representations			P_f	R_f	F_f	P_{on}	R_{on}	F_{on}	P_{onoff}	R_{onoff}	F_{onoff}
STFT:											
N_w											
1024			90.09	87.42	87.73	86.20	76.64	78.74	61.95	56.83	58.09
2048			**90.36**	**90.36**	**89.46**	**89.51**	**77.76**	**80.99**	**66.24**	**59.86**	**61.73**
Mel Spectrogram:											
N_w	N_m										
1024	192		88.41	87.32	86.88	83.13	75.81	76.89	59.87	56.51	56.89
2048	128		90.90	87.32	88.18	85.94	78.32	79.73	62.54	58.85	59.53
2048	192		90.77	85.72	87.20	85.70	76.47	78.49	62.71	58.02	59.10
2048	256		**91.60**	**88.19**	**88.98**	**90.48**	**78.65**	**82.12**	**67.38**	**60.80**	**62.95**
CQT:											
N_b	N_o										
12	7		90.73	88.81	88.91	88.41	77.96	80.76	65.57	60.11	61.68
12	8		90.92	88.50	88.69	89.48	78.00	81.33	66.02	59.70	61.69
24	8		93.02	90.35	90.91	92.67	80.72	84.39	70.79	63.86	66.15
36	8		93.39	90.67	91.27	92.99	80.91	84.71	70.86	63.94	66.29
48	8		**93.89**	90.44	91.44	**93.45**	81.31	85.14	**72.20**	65.15	**67.56**
60	8		93.79	**91.21**	**91.85**	93.25	**81.96**	**85.43**	71.82	**65.18**	67.40
HCQT:											
N_b	N_o	N_h									
36	5	4	91.47	89.43	89.76	91.43	79.79	83.19	67.76	61.55	63.52
60	5	4	91.85	88.96	89.55	90.88	78.88	82.48	66.79	60.47	62.57
60	6	4	92.17	89.88	90.24	90.89	79.55	82.74	67.07	61.11	62.93
60	6	5	**92.97**	**90.49**	**90.95**	**91.81**	**81.06**	**84.14**	**69.27**	**63.53**	**65.31**
60	6	6	91.60	89.03	89.43	88.67	79.48	81.68	64.49	59.93	61.09
VQT:											
N_b	N_o	γ									
36	7	10	92.87	90.64	91.01	92.54	80.60	84.24	70.75	63.87	66.17
60	7	10	93.22	90.69	91.14	92.33	80.37	83.94	71.41	64.49	66.71
60	8	10	94.00	90.93	91.75	**93.94**	82.08	**85.76**	72.63	65.53	67.94
60	8	20	**94.22**	**91.04**	**91.93**	93.81	**82.11**	85.70	**73.07**	**66.25**	**68.53**
60	8	30	94.15	91.01	91.85	93.91	82.03	85.70	73.00	66.05	68.37

size of 8, and inference time/memory is for predicting scores for the whole test set. Score post-processing time is not included in the measurement. We notice that the Lilypond representation inference time is larger than the duration of the music recording, and the Reshaped representation inference time is smaller than the duration of the music recording.

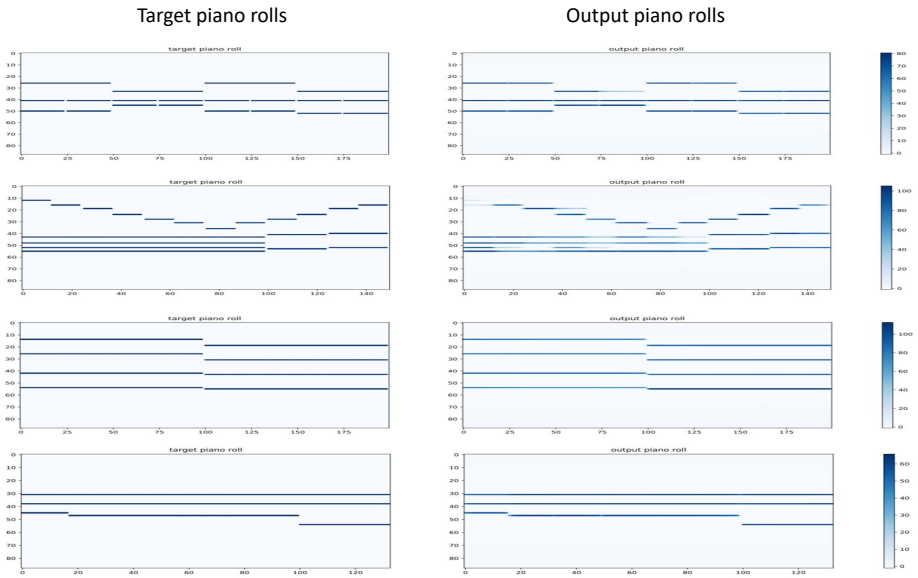

Figure 1: Example velocity-valued piano roll targets and outputs from model with the best input representation. Left column: ground truth piano rolls, Right column: transcribed piano rolls.

Table 2: Time and space used in training (or inference) for the Score-only model with LilyPond representation and Reshaped representation.

Score representation	Time	Memory (MB)
LilyPond	$1194m47s$ ($682m27s$)	10817 (2555)
Reshaped	$156m19s$ ($112m39s$)	5043 (1677)

References

[1] Mert Bay, Andreas F. Ehmann, and J. Stephen Downie. Evaluation of multiple-F0 estimation and tracking systems. In *ISMIR, International Society for Music Information Retrieval Conference*, pages 315–320, 2009.

[2] Yoav Goldberg. *Neural Network Methods for Natural Language Processing*. Morgan & Claypool, 2017.

[3] Andrew Mcleod and Mark Steedman. Evaluating Automatic Polyphonic Music Transcription, 2018.