GCT634: Musical Applications of Machine Learning **Name:** *Vanessa Tan - 20225640*
HW3: Automatic Polyphonic Piano Transcription
November 9, 2022

1 Introduction

Automatic polyphonic piano transcription is the process of converting an audio file into a symbolic form like a musical score or a piano roll [1, 2]. There are many factors to consider when transcribing a piano piece such as pitch, tempo, sustain, dynamics, and many more. Most recent methods in automating polyphonic piano transcription are using deep learning methods particularly recurrent networks since they are great for sequential data. This task explores using recurrent networks in polyphonic music transcription in the frame-level and note-level. Frame-level transcription estimates the pitch that are simultaneously present in each time frame [3]. On the other hand, note-level transcription not only estimates the pitch but also the pitch estimates over time [3].

2 Discussion and Analysis

The different model architectures explored in this experiment are the following: CNN model, Bi-LSTM model, CRNN model (a combination of CNN and LSTM units), and ONF model (a combination of CNN and RNN units with an interconnection between the onsets and frames). The following quantitative metrics were used to evaluate the models: frame F1, onset F1, note F1, note overlap, note F1 (with offset), and note overlap (with offset). As shown in the table below, results show that the best algorithm is the CNN-based model while the worst algorithm is the LSTM-based model.

Table 1: Quantitative Results

Algo	Frame F1	On F1	Note F1	Note Ovrlp	Note F1 (w/ offset)	Note Ovrlp (w/ offset)
CNN	**0.745**	**0.750**	**0.891**	**0.631**	**0.509**	**0.880**
LSTM	0.615	0.490	0.601	0.460	0.248	0.790
CRNN	0.642	0.702	0.814	0.587	0.450	0.872
ONF	0.523	0.669	0.795	0.477	0.344	0.820

As shown in the figures below, the CNN-based model can detect the prominent frames but it cannot detect the sustain of the left hand when the right hand is starting to play a melody. Another observation is that CNN-based model take longer to train than RNN-based models but it converges faster. The CNN-based model also predicts more noise in the frame and onset level than the other models. The disadvantage of RNN-based models is that not all the onsets are predicted correctly particularly in the low frequencies. Another observation that can be seen

from the results is these algorithms are good in detecting frames in the mid-frequency levels. The notes being predicted are also consistent in the frame and onset level. One advantage in using the combination of CNN and RNN units is that they can detect the sustained notes being played.

For future work, a Chroma based input feature can be explored since the focus of piano transcriptions is to predict the correct pitches from the music. Also other RNN models like GRU and transformers could be used to improve the performance of the algorithm. Other metrics could also be created since the sustained note errors did not really reflect in the quantitative metrics used in this experiment.

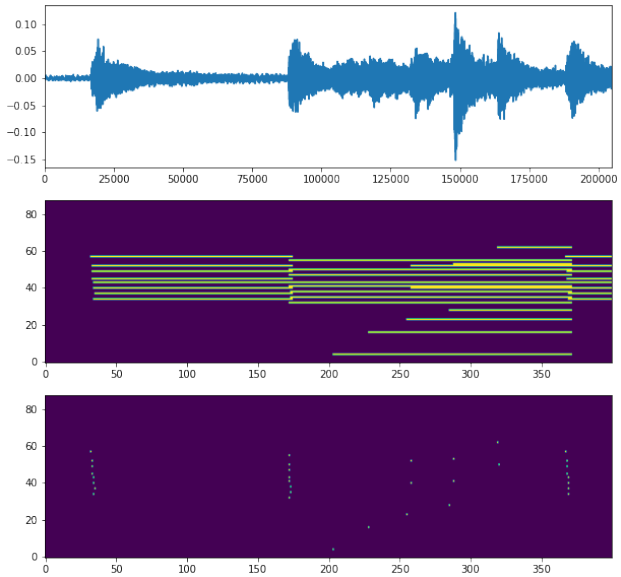

Figure 1: Visualization of Ground Truth Sample

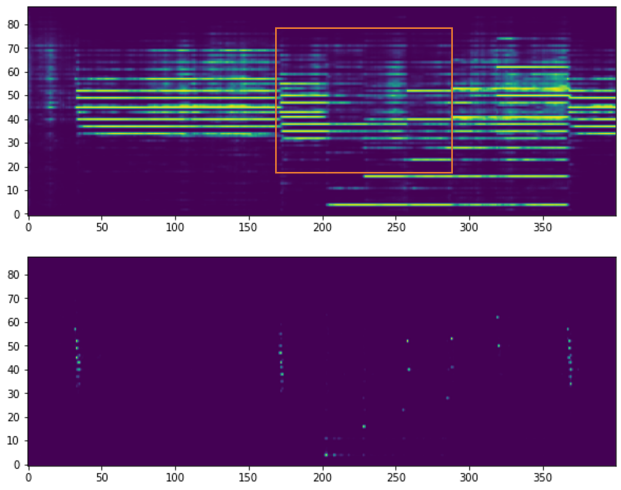

Figure 2: Visualization of CNN Sample

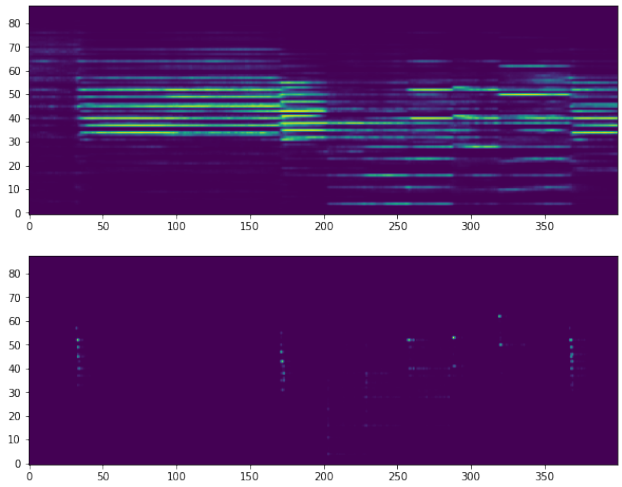

Figure 3: Visualization of Bi-LSTM Sample

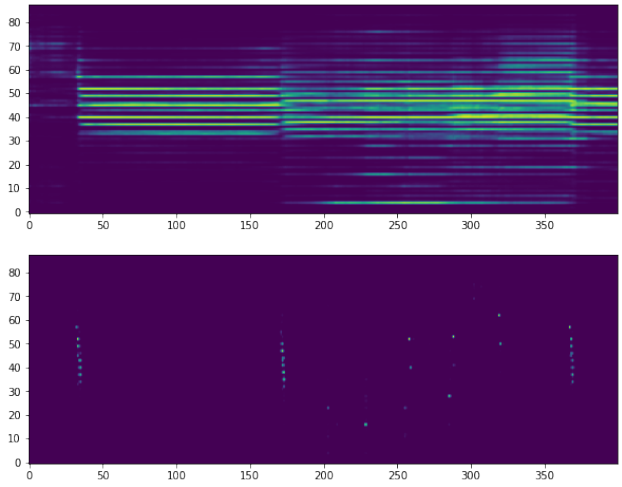

Figure 4: Visualization of CRNN Sample

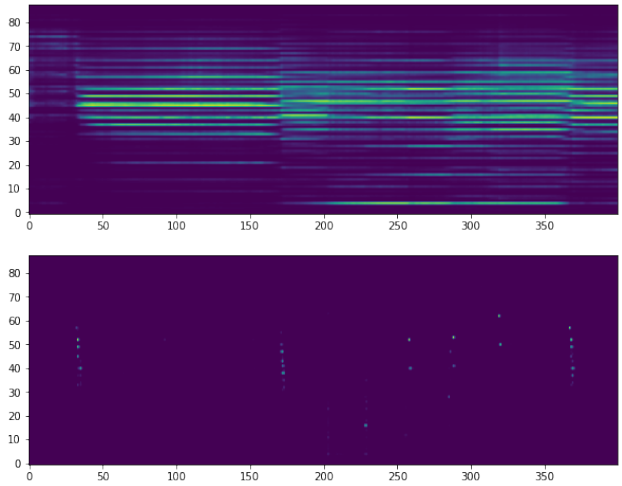

Figure 5: Visualization of ONF Sample

3 Conclusion

This experiment demonstrates the polyphonic piano transcription using CNN and RNN units. Quantitative results show that the CNN-based model is the best algorithm while the RNN-based model is the worst algorithm. Qualitative results show that the combination of CNN and RNN units have an advantage. They can better predict sustained notes when a melody is being played which is most of the time being played in a piano piece. Other metrics and model architectures could be explored to improve the performance of automatic polyphonic piano transcriptions.

References

[1] C. Hawthorne, E. Elsen, J. Song, A. Roberts, I. Simon, C. Raffel, J. Engel, S. Oore, and D. Eck, "Onsets and frames: Dual-objective piano transcription," 2017.

[2] T. Kwon, D. Jeong, and J. Nam, "Polyphonic piano transcription using autoregressive multi-state note model," 2020.

[3] E. Benetos, S. Dixon, Z. Duan, and S. Ewert, "Automatic music transcription: An overview," *IEEE Signal Processing Magazine*, vol. 36, no. 1, pp. 20–30, 2019.