Viktor Pavlov 230335TAF

1 Problem Statement

In the domain of algorithmic music composition, different strategies can lead to a wide range of outputs in terms of quality and the style of the results. The goal of this study is to explore and compare the musical outputs of two types of models: Music Transformer [2] model and Recurrent Neural Network (RNN) models. The Music Transformer model extends the classic Transformer model with relative attention mechanism, which modulates attention based on how far apart two tokes are [5]. The relative self-attention mechanism allows the model to generate outputs with long-term coherence that extends beyond the length of the training examples. In comparison to the Transformer model, RNNs, including Melody RNN and Performance RNN [4], are simpler models that can be used to generate monophonic melodies and performance-like polyphonic sequences respectively.

2 Music Transformer

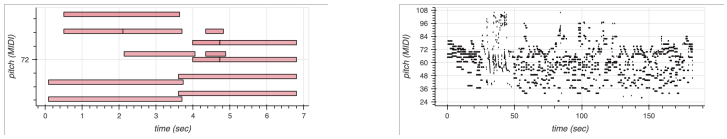

Figure 1: Continuation (right) generated by the Unconditional Transformer model based on primer (left).

In the course of this study, I utilized two Music Transformer models from Google's Magenta library [6]: Unconditional and Score Conditioned. The Unconditional model, although is capable of generating continuations from a given primer melody (as can be seen in Figure 1), can also generate music without an initial seed, relying on its understanding of musical patterns and structure. On the other hand, the Score Conditioned model works in a sequence-to-sequence manner by generating new interpretations of the provided score. This includes capabilities like generating new accompaniment for a given melody, similarly to adding a new dimension to the input (illustrated in Figure 2). Both models were trained on the MAESTRO dataset [1]. The MAESTRO (MIDI and Audio Edited for Synchronous TRacks and Organization) is a dataset composed of about 200 hours of virtuosic piano performances captured with fine alignment ($\sim$ 3ms) between note labels and audio waveforms. Additionally, the models use the score-to-performance (Score2Perf [7]) encodings that enrich the MIDI [3] grid-like representations of music with slight variations in timing and velocity in order to achieve a more human-like sound.

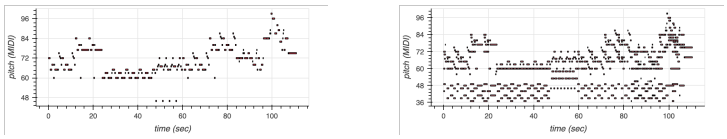

Figure 2: Accompaniment (right) generated by the Score Conditioned Transformer model based on melody (left).

3 RNNs

For comparison to the Transformer model, Recurrent Neural Network (RNN) models were utilized from the Magenta library: Melody RNN and Performance RNN. The Melody RNN is a model specifically designed to handle monophonic melodies, while the Performance RNN model is designed for polyphonic piano music with expressive timing and dynamic. Both models have a LSTM-based architecture and were trained on large dataset of MIDI files, allowing them to learn a wide range of musical styles and sturctures.

4 Model Comparison

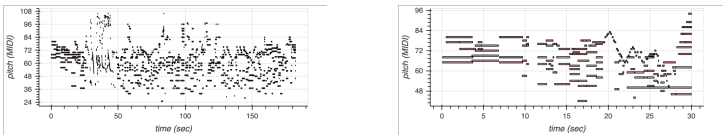

Figure 3: Transformer continuation (left) and RNN continuation (right).

In order to compare the models I used the following criteria: musical quality, efficiency and flexibility.

Musical Quality The melody continuation task was used to evaluate the musical quality. The Transformer model demonstrated superior performance in terms of maintaining musical coherence when compared to the RNN. However, given that both models were trained on virtuosic piano music, the generated sequences included elements of improvisation, some of which seemed out of context.

Efficiency In terms of performance the RNN model was significantly faster, as can be seen in Table 1. The Transformer model is known to be computationally heavy due to its self-attention mechanism, which computes pairwise interactions between all input positions. On the other hand, the RNN model is generally faster and less computationally demanding, as it processes the sequence one step at a time. However, it might struggle with long sequences due to the vanishing gradient problem, which can make it difficult for the model to learn long-range dependencies.

Model	**Output Duration (s)**	**Execution (s)**
Transformer	183	$\sim$ 90
RNN	31	2

Table 1: Comparison of models in terms of output duration and execution time.

Flexibility Both types of models offered some flexibility. The RNNs provided options for generating both monophonic and polyphonic sequences. However, what truly stood out was the Transformer's ability to reinterpret a musical score, by generating new harmony for the melody.

References

[1] Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse Engel, and Douglas Eck. Enabling factorized piano music modeling and generation with the MAESTRO dataset. In *International Conference on Learning Representations*, 2019.

[2] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Curtis Hawthorne, Andrew M Dai, Matthew D Hoffman, and Douglas Eck. Music transformer: Generating music with long-term structure. *arXiv preprint arXiv:1809.04281*, 2018.

[3] MIDI Manufacturers Association. Midi 1.0 detailed specification. `https://www.midi.org/specifications-old/item/the-midi-1-0-specification`, 1996.

[4] Sageev Oore, Ian Simon, Sander Dieleman, and Doug Eck. Learning to create piano performances. In *NIPS 2017 Workshop on Machine Learning and Creativity*, 2017.

[5] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations, 2018.

[6] Google Brain Team. Magenta. `https://magenta.tensorflow.org/`, 2023. Accessed: 2023-06-09.

[7] Magenta Team. Magenta score2perf model. `https://github.com/magenta/magenta/tree/main/magenta/models/score2perf`, 2023.