Visualizing Music Self-Attention

Anna Huang[*]
Google Brain
annahuang@google.com

Monica Dinculescu[*]
Google Brain
noms@google.com

Ashish Vaswani
Google Brain
avaswani@google.com

Douglas Eck
Google Brain
deck@google.com

Abstract

Like language, music can be represented as a sequence of discrete symbols that form a hierarchical syntax, with notes being roughly like characters and motifs of notes like words. Unlike text however, music relies heavily on repetition on multiple timescales to build structure and meaning. The Music Transformer has shown compelling results in generating music with structure [3]. In this paper, we introduce a tool for visualizing self-attention on polyphonic music with an interactive pianoroll. We use music transformer as both a descriptive tool and a generative model. For the former, we use it to analyze existing music to see if the resulting self-attention structure corroborates with the musical structure known from music theory. For the latter, we inspect the model's self-attention during generation, in order to understand how past notes affect future ones. We also compare and contrast the attention structure of regular attention to that of relative attention [6, 3], and examine its impact on the resulting generated music. For example, for the JSB Chorales dataset, a model trained with relative attention is more consistent in attending to all the voices in the preceding timestep and the chords before, and at cadences to the beginning of a phrase, allowing it to create an arc. We hope that our analyses will offer more evidence for relative self-attention as a powerful inductive bias for modeling music. We invite the reader to view our video animations of music attention and to interact with the visualizations at `https://storage.googleapis.com/nips-workshop-visualization/index.html`.

1 Introduction

Attention is a cornerstone in neural network architectures. It can be the primary mechanism for constructing a network, such as in the self-attention based Transformer, or serve as a secondary mechanism for connecting parts of a model that would otherwise be far apart or different modalities of varying dimensionalities. Attention also offers us an avenue for visualizing the inner workings of a model, often to illustrate alignments [4]. For example in machine translation, the Transformer uses attention to build up both context and alignment while in the LSTM-based seq2seq models, attention eases the word alignment between source and target sentences. For both types, attention gives points us to where a model is looking when translating [7, 1]. For example in speech recognition, attention aligns different modalities from spectograms to phonemes [2].

In contrast to the above domains, there is less "groundtruth" in what should be attended to in a creative domain such as music. Moreover, in contrast to encoder-decoder models where attention serves as alignment, in language modeling self-attention serves to build context, to retrieve relevant information

[*]Equal contribution.

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

from the past to predict the future. Music theory gives us some insight of the motivic, harmonic, temporal dependencies across a piece, and attention could be a lens in showing their relevance in a generative setting, i.e. does the model have to pay attention to this previous motif to generate the new note? Music Transformer, based on self-attention [7], has been shown to be effective in modeling music, being able to generate sequences with repetition on multiple timescales (motifs and phrases) with long-term coherence [3]. In particular, the use of relative attention improved sample quality and allowed the model generalize beyond lengths observed during training time. Why does relative attention help? More generally, how does the attention structure look like on these models?

In this paper, we introduce a tool for visualizing self-attention on music with an interactive pianoroll. We use Music Transformer as both a descriptive tool and a generative model. For the former, we use it to analyze existing music to see if the resulting self-attention structure corroborates with musical structure known from music theory. For the latter, we inspect the model's self-attention during generation, in order to understand how past notes affect future ones. We explore music attention on two music datasets, JSB Chorales and Piano-e-Competition. The former are Chorale harmonizations, and we see attention keeping track of the harmonic progression and also voice-leading. The latter are virtuosic classical piano music and attention looks back on previous motifs and gestures. We show for JSB Chorales the heads in multihead-attention distribute and focus on different temporal regions.

Moreover, we compare and contrast the attention structure of regular attention to that of relative attention, and examine its impact on the resulting generated music. For example, for the JSB Chorales dataset, a model trained with relative attention is more consistent in attending to all the voices in the preceding timestep and the many chords before, and at cadences to the beginning of a phrase, allowing it to create an arc. In contrast, regular attention often becomes a "local" model only attending to the most recent history, resulting in certain voice repeating the same note for a long duration, perhaps due to overconfidence.

2 Background

2.1 Data representation

We take a language-modeling approach to training generative models for symbolic music. Hence we represent music as a sequence of discrete tokens, with the vocabulary determined by the dataset. The JSB Chorale dataset consists of four-part scored choral music, which can be represented as a pianoroll like representation with rows being pitch and columns being time discretized to sixteenth notes. It is serialized in raster-scan fashion when consumed by a language model. For the Piano-e-Competition dataset we use the performance encoding [5] which consists of a vocabulary of 128 `NOTE_ON` events, 128 `NOTE_OFF`s, 100 `TIME_SHIFT`s allowing for expressive timing at 10ms and 32 `VELOCITY` bins for expressive dynamics.

2.2 (Relative) self-attention in Transformer

The Transformer [7] is a sequence model based primarily on self-attention. Multiple heads are typically used to allow the model to focus on different parts of the history. These are supported by first splitting the queries Q, keys K, and values V into h parts on the depth d dimension.

Equation 1 shows the scaled dot-product attention for a single head. Regular attention consists of only the $Q^h K^{h\top}$ term, while relative attention adds S^{rel} to modulate the attention logits based on pairwise distances between queries and keys.

$$Z^h = \text{Attention}(Q^h, K^h, V^h) = \text{Softmax}\left(\frac{Q^h K^{h\top} + S^{rel}}{\sqrt{D_h}}\right) V^h. \tag{1}$$

We adopt $S^{rel} = \text{Skew}(Q^h E^{h\top})$ as in [3], where E^h are learned embeddings for every possible pairwise distance. The attention outputs for each head are concatenated and linearly transformed to get Z, a L by D dimensional matrix, where L is the length of the input sequence.

3 Our music attention visualization tool

Figure 1 shows a full-view of our visualization tool for exploring self-attention [2]. The arcs, inspired by [8], connect the current query (highlighted by the pink playhead) to earlier parts of the piece. Each head bears a different color, and the thickness of the lines give the attention weights. The user can choose to see a subset of the attention arcs either by specifying the top n number of arcs or by specifying a threshold at which attention weights lower then that would not be shown. Our tool also supports animation, which allows us to inspect if a certain phenomena is consistent throughout a piece, and not just for certain timesteps.

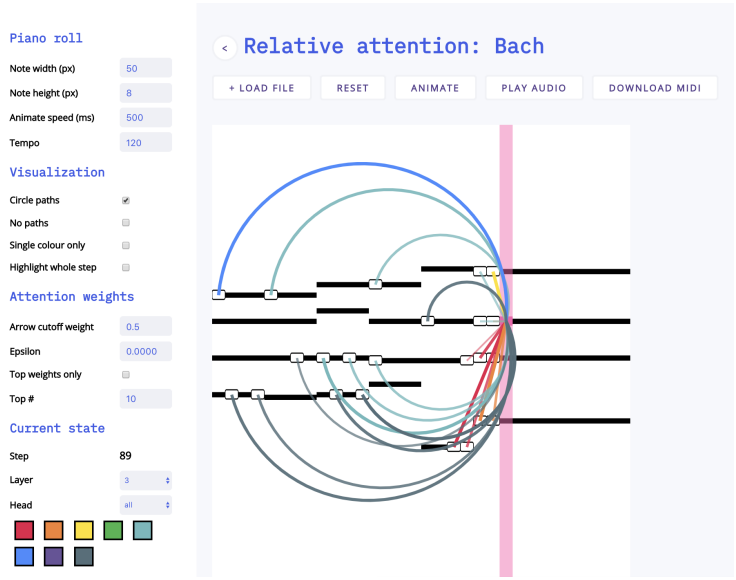

Figure 1: Screenshot of our music attention visualization tool, showing multi-head self-attention on a generated pianoroll.

4 Preliminary observations

4.1 Heads distribute in time

Figure 1 shows that some heads focus on the immediate past, some further back, nicely distributed in time. This maybe due to relative attention explicitly modulating attention based on pairwise distance.

4.2 Bottom dense, top sparse

The left shows how on the bottom layer the attention is dense, while the right shows on the top layer each position is already a summary and hence the model only needs to attend to less positions.

Figure 2: Self-attention is dense in the bottom layer (left) while sparse in the top layer (right).

[2] All the figures have an interactive version at
`https://storage.googleapis.com/nips-workshop-visualization/index.html`

4.3 Regular attention collapses to a local model

When trained on JSB Chorales, regular attention failed to align the voices, causing one voice to repeat the same note (left on Figure 3), while relative attention generated samples with musical phrasing.

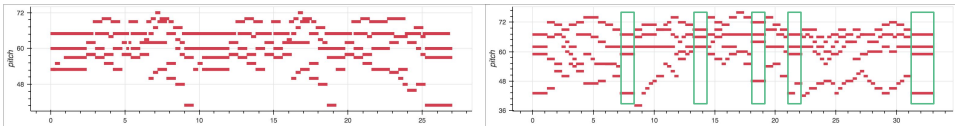

Figure 3: Unconditioned samples from Transformer without (left) and with (right) relative self-attention. Green vertical boxes indicate the endings of (sub)phrases where cadences are held.

To compare, we use relative attention (pink) and regular attention (green) to analyze the same JSB Chorale, by feeding the piece through the models and recording their attention weights. Figure 4 shows a drastic difference in how regular attention only focuses on the immediate past and the beginning of the piece, while relative attention attends to the entire passage.

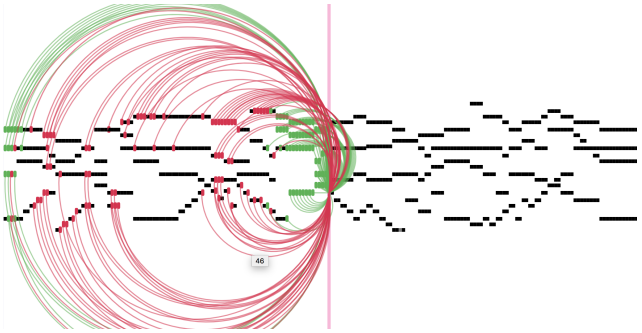

Figure 4: Analysis of an existing JSB Chorale: regular attention (green) only attends locally plus to the beginning of the piece, while relative attention (pink) keeps track of a whole phrase.

4.4 Attending to motifs while generating piano performances

Figure 5 shows a sample generated by Transformer with relative attention trained on the Piano-e-Competition dataset. The top shows a passage with right-hand "triangular" motifs and the model attends to the runs to learn the scale and also peaks to know when to change directions. The bottom shows the same passage with the query being on the left-hand, and the attention focuses more on the left-hand chords and also the right-hand when they coincide with the left-hand.

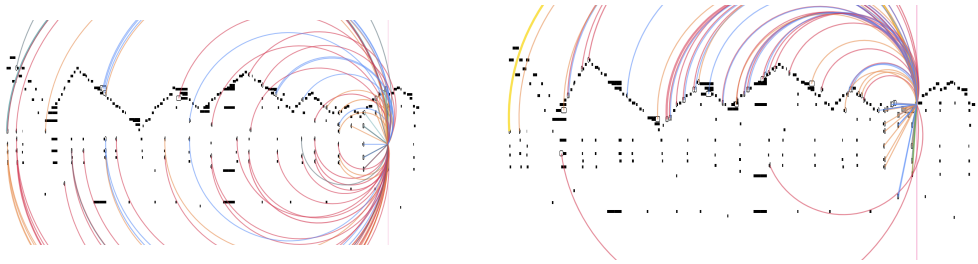

Figure 5: On the left panel, the query is a left-hand note and attention is more focused on the bottom half compared to the right which is right-hand note, with more attention on the top half.

5 Conclusion

We presented a visualization tool for seeing and exploring music self-attention in context of music sequences. We have shown some preliminary observations and we hope this it the beginning to furthering our understanding in how these models learn to generate music.

References

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[2] Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. Attention-based models for speech recognition. In *Advances in neural information processing systems*, pages 577–585, 2015.

[3] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M. Dai, Matthew D. Hoffman, Monica Dinculescu, and Douglas Eck. Music transformer. *arXiv preprint arXiv:1809.04281*, 2018.

[4] Chris Olah and Shan Carter. Attention and augmented recurrent neural networks. *Distill*, 2016.

[5] Sageev Oore, Ian Simon, Sander Dieleman, Douglas Eck, and Karen Simonyan. This time with feeling: Learning expressive musical performance. *arXiv preprint arXiv:1808.03715*, 2018.

[6] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, volume 2, pages 464–468, 2018.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010, 2017.

[8] Martin Wattenberg. Arc diagrams: Visualizing structure in strings. In *Information Visualization, 2002. INFOVIS 2002. IEEE Symposium on*, pages 110–116. IEEE, 2002.