[Paper Review]
Music Transformer Generating Music With Long-Term Structure

Paul Jason Mello
Department of Computer Science and Engineering
University of Nevada, Reno
pmello@unr.edu

Abstract

”Music relies heavily on repetition to build structure and meaning. Self-reference occurs on multiple timescales, from motifs to phrases to reusing of entire sections of music, such as in pieces with ABA structure. The Transformer (Vaswani et al., 2017)[5], a sequence model based on self-attention, has achieved compelling results in many generation tasks that require maintaining long-range coherence. This suggests that self-attention might also be well-suited to modeling music. In musical composition and performance, however, relative timing is critically important. Existing approaches for representing relative positional information in the Transformer modulate attention based on pairwise distance (Shaw et al., 2018) [4]. This is impractical for long sequences such as musical compositions because their memory complexity for intermediate relative information is quadratic in the sequence length. We propose an algorithm that reduces their intermediate memory requirement to linear in the sequence length. This enables us to demonstrate that a Transformer with our modified relative attention mechanism can generate minute-long compositions (thousands of steps) with compelling structure, generate continuations that coherently elaborate on a given motif, and in a seq2seq setup generate accompaniments conditioned on melodies1. We evaluate the Transformer with our relative attention mechanism on two datasets, JSB Chorales and Maestro, and obtain state-of-the-art results on the latter.” [3]

1 Summary

The authors of this paper improve on prior transformer architecture choices for the task of music generation through long-term sequencing. They are motivated by the application of attention for music and describe the challenges they overcame. These challenges are defined by significant memory overhead and the complex interdependencies between music motifs, including pitch, timing, and more. To overcome these obstacles they propose a novel attention mechanism described as ”relative positional self-attention”, which solves both issues by converting the transformer architecture from an absolute-by-relative indexing to an absolute-by-absolute indexing. They achieve this transformation through a set of heuristic matrix operations designed to embed positional information about distances between musical motifs. Through this process to improve matrix efficiency and long-term similarity, they achieve a reduction in memory requirements, going from 8.5 GB to 4.2 MB, per layer.

2 Main Contributions

Each of the three contributions they propose allows for the creation of a highly effective music generator architecture. Prior work has focused on LSTMs, which have shown success in sequence modelling, but more novel approaches have used style transfer, and CNNs trained as GANs, to achieve similar results in music generation. Despite these interesting designs, no work had yet to utilize the transformer architecture in their work.

2.1 Key Contributions

- Firstly, this is the first paper to successfully apply transformer architectures to the task of music generation.
- Secondly, they develop a relative positional self-attention mechanism. This is crucial to allow for the continuation of music through motifs and generate a similar style to the initial input sequence. Their implemented attention mechanism also has the added benefit of generalizing and generating sequences beyond the initial input length that are coherent and build off the prior musical motifs.
- Finally, the authors notice the lack of proper positional encoding architectures in prior work in self-attention mechanisms, like that of Shaw et al. [4]. Shaw also has a significant space complexity which the authors present a solution for. They reduce the memory requirements of attention from $\mathcal{O}(L^2D)$ to $\mathcal{O}(LD)$, where L is the sequence length, and D is the dimensionality. This results in a memory reduction per layer from 8.5 GB to 4.2 MB, when $L = 2056$ and $D = 512$.

2.2 Innovative Aspects

The main innovations of this work is the matrix operations used to encode positional awareness into the transformer architecture and reduce memory requirements at the same time. They call this approach to attention relative positional self-attention. In doing so the authors generated a linear time model which is scalable for long-term sequences with efficient long-term structure modeling.

- The authors replace the standard absolute positional encoding from the transformer architecture. In music generation, the timing and pitch of notes significantly effect the overall quality and perception of the music and rhythm. Traditional absolute positional encoding sequences utilize a sinusoidal positional embedding to capture this, but this is insufficient. This approach fails to model complex interdependence's as seen in music. With the introduction of relative positional self-attention, the timing and pitch are relatively encoded, allowing the model to learn the relationships between any two notes in a more flexible format. These relationships are captured through pairwise embeddings that represent the distance between any two tokens in the sequence.
- The memory complexity is reduced through precise matrix operations and a concept called "skewing". Firstly, the authors avoid constructing a full $\mathcal{L} \times \mathcal{L} \times \mathcal{D}$ tensor by taking the relative difference $r = j - i$ between any query at position i and key at position j, thus creating an efficient embedding we denote as $E_r[r]$. We then compute the product of these operations between the query vector Q_i and the relative positional embedding, resulting in a matrix S_{rel}. The skewing mechanism works by first adding a left column of padding to the matrix $QE_r^\top$. This padded matrix is then reshaped into an $(L + 1) \times L$ matrix. Finally, the matrix is sliced to remove the last row, keeping only the valid information. This results in the relative logits aligning correctly with the query-key pairs, enabling efficient computation of relative positional attention and the generation of long term sequences.

3 Strengths and Weaknesses

The strengths of this paper lies in the optimization of prior architectures and application to a novel data structure. This paper achieves its goals in a general way leading to great long-term sequencing results. However, their model evaluation approach has a few weakness including a lack of training on more datasets which may offer better insights and generalization.

3.1 Strengths

The primary strength of this work is the reduced space complexity and efficient positional embedding. As described in 2.2, the memory requirements are reduced from being squared to linear. The architecture design also utilizes general matrix operations to adjust the transformer from an absolute-by-relative to an absolute-by-absolute model which allows for an expansion to a diverse set of applications beyond the task of music generation. The effectiveness of these strengths are easily identifiable by the output quality of the model, where the generated music is able to generalize beyond the initial input sequence. Empirically they demonstrate final results showing a significant improvement over prior related works. The authors then compare their model to other music tasks demonstrating the flexibility of their approach to unconditioned music generation, priming, and accompaniment generation.

3.2 Weaknesses

While the paper is sound and particularly thorough, there is one main weakness. Particularly, they explore only classical music datasets. Their use of classical datasets are akin to having a very clean and focused dataset. It would be interesting to see how this model fairs in other, more free-flowing, music genres like jazz or rock, in training, inference, and even evaluation to unseen music motifs. These music genres and flows may offer more complex, or rather non-standard patterns, and would show the capabilities of the model to generalize beyond the data, and beyond similarities in one music genre.

3.3 Areas of Improvements

The only significant area of improvements to this paper would be to discuss the time complexity of their upgraded attention mechanism. While they are able to significantly reduce the space complexity, they do not mention the time complexity beyond the initial transformer, which takes quadratic time. The assumption would then be their model remains quadratic in time. However, the lack of clarity regarding the time complexity leaves much to the imagination and would not have been a challenging experiment to run and add to the work. I am particularly picky on this as, in my opinion, you can not discuss space complexity without discussing time complexity as well.

4 Discussion

If the final output quality of the model is representative of the quality of the novel contributions, then this paper has contributed substantially. The novel application of a transformer to the task of music generation, and the advancements in space complexity, which inherently handles pair-wise encoding for long-term sequence dependencies, is a novel idea which proves to work succinctly together. It demonstrates both the effectiveness of transformer architectures on long-term sequence outputs and the application to complex data like music. It would be wonderful to revisit this paper today with our novel improvements to the attention mechanism and see the quality of the generated music. Particularly attention mechanisms in the same problem domain such as, Performer [2], Linformer [6], and Longformer [1] which directly proceeded the initial transformer model.

5 Conclusion

The authors of this work introduce improvements to relative positional self-attention mechanisms by tailoring the architecture for applications in music generation. They recognize points of improvements to the transformer architecture and integrate these improvements into the design by capturing the relationships between features in a memory efficient manner. The authors saturate every intermediate layer with this transformer architecture and demonstrate a significant improvement over the state of the art. Prior works leveraged LSTMs to solve music generation beyond the initial sequence length, but found little success when compared to the transformer approach. This approach demonstrates that, with the right tweaks, foundational attention is capable of symbolic music generation that match pitch and tone while also completing melodies and matching motifs.

References

[1] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer, 2020.

[2] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy Colwell, and Adrian Weller. Rethinking attention with performers, 2022.

[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, 2018.

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

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[6] Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity, 2020.