Exploring the Realm of Automated Music Composition with Models based on Transformer and its Variants

Shuanglong Zhu *

Department of Computing, Queen's University, Kingston, Canada

* Corresponding Author Email: 19sz62@queensu.ca

Abstract. The Transformer architecture and its modelling variants have great potential in the music domain. Traditional approaches, such as rule-based systems and RNNs, have limitations in capturing the complex temporal dependencies and hierarchical structures inherent in music. With the introduction of its self-concern mechanism, the Transformer can effectively address this problem by capturing remote dependencies and allowing parallel processing of sequences. Transformer-based model variants such as Music Transformers, MuseNet and Jukebox have demonstrated that they can generate high-quality, varied, and stylistically rich compositions. Despite the success of these models, some challenges still need to be solved, such as high computational requirements and limited control over musical style and emotion. Possible future research directions include optimizing computational efficiency, enhancing stylistic and emotional control, and developing hybrid models that combine other models with Transformer. This article provides a comprehensive overview of the current state of research and potential and future applications of Transformer for music generation.

Keywords: Transformer, Self-Attention Mechanism, Relative Positional Encoding, Sparse Attention, Variational Autoencoders (VAEs).

1. Introduction

The application of deep learning techniques has significantly impacted the field of music. Traditional automated compositional methods typically rely on rule-based systems or Markov chains, which are limited in capturing the complex temporal structure inherent in music. The advent of neural networks has brought about a paradigm shift, allowing for generating more coherent and musically rich sequences.

Various deep learning models, in particular Recurrent Neural Networks (RNN) and their variants, such as Long Short-Term Memory Networks (LSTM) and Gated Recurrent Units (GRU), have been used for music generation [1]. These models have shown promise in capturing long-term dependencies and generating more complex musical sequences than traditional methods. Notable works include using LSTMs for blues improvisation and generating Bach-style choral harmonies. However, these RNN-based models still face significant challenges, such as the difficulty of capturing long-term dependencies and the gradient vanishing problem, which limits their effectiveness for longer sequences.

The Transformer proposed by Vaswani et al. addresses these challenges through its self-attentive mechanism, which allows long-distance dependencies to be captured efficiently by processing sequences in parallel [2]. This mechanism allows each note to generate other notes in the entire musical sequence, thus integrating global information and producing more coherent and complex compositions. The Transformer model and its variants (e.g., Music Transformer, MuseNet, and Jukebox) generate high-quality, varied, and stylistically rich musical compositions. This paper explores the impact and application of the Transformer model and its variants in music generation, emphasizing their ability to capture complex musical structures and generate high-quality compositions. The main goals of this paper are to provide an overview of the different Transformer architectures and their key components, to review current Transformer-based models for music generation, to analyze the impact of these models on the quality and diversity of the music generated, and to discuss the potential future directions and applications of Transformer models in music

composition. These aspects clearly illustrate the potential of Transformers in the field of music generation and promote further research and development in this area.

2. Main Body

2.1. The development of automated music composition

The journey to automated music composition began with rule-based systems and statistical methods such as Markov chains. While innovative, limiting these approaches is their inability to capture the complex temporal dependencies and hierarchical structure in music. Early neural network models, including simple feed-forward and early recurrent neural networks (RNNs), attempted to address these limitations but faced significant challenges in generating coherent and musically pleasing compositions.

However, the development of Recurrent Neural Networks (RNN) and their variants, such as Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRUs), has brought significant advances to the field of music generation [1]. LSTMs, in particular, are well suited for melody and harmony generation tasks due to their expertise in capturing long-term dependencies in sequences. Notable works include Eck and Schmidhuber, who explored the use of LSTMs for blues improvisation [1], and Hadjeres and Pachet, who introduced DeepBach, a model for generating Bach-style choral harmonies using LSTMs [3]. Despite their success, RNN-based models still face problems such as vanishing gradients and limited parallelization capabilities, which make them less efficient for longer sequences.

The Transformer architecture proposed by Vaswani et al. revolutionized sequence modelling by replacing loops with a self-attentive mechanism [2]. The Transformer captures long-distance dependencies and allows for parallel sequence processing, significantly improving training efficiency. Huang et al. introduced Music Transformer, a specialized music generation. By incorporating relative self-attention, Music Transformer can efficiently capture the hierarchical structure of music, generating compositions with longer-term coherence and more complex structures than RNN-based models [4]. Payne presented MuseNet, an extension of the Transformer model capable of generating various styles, from classical to contemporary genres. MuseNet demonstrates the potential of Transformer to create diverse and stylistically rich compositions [5]. Dhariwal et al. introduced JukeBox, a Transformer-based model for generating high-fidelity music with lyrics [6]—transformers to generate coherent and context-rich musical compositions.

Comparative studies have shown that Transformer-based models outperform traditional RNN-based models in terms of coherence, diversity, and ability to capture long-term dependencies. For example, Huang et al. demonstrated that Music Transformers can generate musically more pleasing and structurally coherent compositions than LSTM-based models [7]. Thus, Transformers' success in music generation opens up new avenues for research and applications. Future directions include exploring hybrid models that combine the strengths of RNNs and Transformers, developing more efficient training techniques, and extending the application of Transformers to other areas of music creation, such as real-time improvisation and interactive music systems.

The Transformer architecture has found extensive application in automated music composition. Its core components make it particularly suited for music generation tasks. Below is a detailed explanation of the Transformer architecture and its various variants in automated music composition.

2.2. Transformer and their variant

2.2.1. Transformer

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \quad (1)$$

The self-attention mechanism of the Transformer model allows each note to be generated based on other notes in the entire musical sequence, thus enabling the integration of global information. This mechanism, described by the formula (1), where Q, K, and V represent the query, key, and value matrices, captures complex musical dependencies and hierarchies, effectively modelling long-distance dependencies in the sequence [2]. For example, in an extended musical sequence, early chords or melodic fragments can affect subsequent musical fragments, and the self-attention mechanism effectively captures such long-distance dependencies. In music generation tasks, this mechanism allows the model to refer to the entire musical sequence when generating new notes, resulting in more coherent and complex musical compositions.

2.2.2. Musenet

Building on Transformer's basic functionality, MuseNet introduces additional flexibility and extensibility. MuseNet is a Transformer-based model that generates music in various classical and contemporary styles. Unlike traditional models, MuseNet is not explicitly programmed using music theory; instead, it learns harmony, rhythm, and style patterns by predicting the next token in hundreds of thousands of MIDI files. GPT-2 is a large-scale Transformer model trained to predict the next marker in a sequence, whether audio or text [5].

In addition, MuseNet uses embeddings to represent various musical features, such as pitch, volume, and instrument information and combines them into a single token. It helps the model understand and generate complex musical structures. In addition, MuseNet employs a sparse transformer that optimizes the architecture to handle long sequences efficiently. The model uses a 72-layer network and 24 attention heads that focus on the context of 4096 tokens, which helps MuseNet capture long-term structures in music, such as remembering themes and maintaining coherence in extended sequences [5]. However, the main limitation of MuseNet is that it relies on MIDI files rather than raw audio, which limits its ability to produce high-fidelity music with nuanced vocal and instrumental timbres.

2.2.3. Jukebox

To address the limitations of MuseNet, Dhariwal et al. proposed Jukebox, a generative model that specializes in creating high-fidelity music in the raw audio domain. It uses multi-scale VQ-VAE to compress raw audio into discrete codes, then modeled using an autoregressive transformer. This approach ensures the necessary musical information while reducing computational complexity, allowing for generating high-fidelity, and diverse songs with several minutes of coherence [6]. Jukebox employs a sparse Transformer to efficiently generate long audio sequences that can be adapted to specific artists and genres to influence musical and vocal styles and even control singing based on inconsistent lyrics. The model effectively addresses the challenge of long-distance dependencies in raw audio, allowing for generating music with complex structures and multiple instruments [6]. Despite Jukebox's advances, its complexity and computational demands can be a drawback, requiring significant resources for training and generation.

2.2.4. Music Transformer

While MuseNet and Jukebox demonstrate substantial diversity and stylistic richness capabilities, they have limitations in capturing long-term structure and complexity in music. In contrast, Music Transformer, proposed by Huang et al., focuses on utilizing self-attentive mechanisms to generate music with long-term structure. Music Transformer has significant advantages in capturing musical structure and generating long-term coherent music, especially in handling relative time and pitch relationships. Unlike recurrent neural networks, which rely on sequential processing and fixed-size memory states, the self-attention mechanism allows Music Transformer to refer to any part of a previously generated sequence at each step, thereby capturing musical elements that recur at various levels, such as themes, phrases and repetitive segments. This capability is particularly suited to music, where repetition and variation in long sequences are critical for coherence and expressiveness [4].

$$RelativeAttention(Q, K, V, S) = Softmax(\frac{QK^T + S^{rel}}{\sqrt{D_h}})V \quad (2)$$

To deal with the crucial relative time and pitch in musical compositions, Music Transformer also extends the Relative Self-Attention mechanism to capture these pairwise relationships efficiently, addressing key aspects of time and pitch. The formula (2) represents the relative self-attention, where Q, K, and V represent the query, key, value matrices and the relative position encoding [8]. Relative self-attention represents relative positional relationships, thus more effectively capturing structural information in music such as repeated themes, call and response, scale and gauge and arpeggio patterns [4]. In music generation, relative self-attention better captures repetition and variation in musical sequences, resulting in more complex and coherent musical compositions. In addition, the music transformer employs a memory-efficient implementation of relative attention from $O(L^2D)$ to O(LD), reducing intermediate memory requirements and enabling the Transformer to train long sequences and produce structurally coherent, long-term musical segments [4]. However, while Music Transformer excels at working with hierarchies and producing complex compositions, it operates primarily in pre-segmented musical sequences, limiting its ability to work with raw audio data.

2.2.5. Transformer-XL

While they made significant progress with the Transformer and other variants, there is potential for further development of Transformer-XL, a variant of the Transformer model that focuses on capturing long-term dependencies, as proposed by Dai et al. [9]. Transformer-XL extends the length of contexts that can process beyond the capabilities of the original Transformer and RNNs by combining a segment-level recursion mechanism with a novel positional encoding scheme. This architecture allows the model to maintain information flow across segments and address context fragmentation, ensuring global consistency between the original and growing music sequences.

Transformer-XL's ability to reuse hidden states from previous segments as memory enables it to capture long-term dependencies without disrupting temporal coherence. It makes it possible to generate coherent long-form musical compositions that maintain a consistent structure and style over time. The introduction of relative positional coding further improves the model's ability to generalize to longer sequences during evaluation, thus increasing its effectiveness in music generation tasks. These features make Transformer-XL highly effective in automated music composition, allowing for the creation of high-quality, coherent and diverse musical works.

2.3. The improvement and future of Transformer

2.3.1. General Challenges in Automated Music Composition

Despite significant advances in automated music composition, several inherent challenges remain. Firstly, automated systems often struggle to replicate the depth of human emotion and creativity inherent in music. While algorithms can generate technically proficient compositions, they frequently need more emotional nuance and inventive spark that characterize human-created music. The challenge lies in imbuing AI-generated music with the same emotional resonance and originality a human composer can provide.

Music involves notes, rhythms, and intricate structures, including harmonies, counterpoints, and thematic developments. Traditional rule-based systems and even advanced neural networks sometimes need help to capture these complex structures, leading to compositions that may sound mechanical or lack cohesion over extended pieces. Additionally, while current models can generate music in various styles, they often need to catch up in accurately capturing the nuances of different genres and adapting to the stylistic requirements of diverse musical forms—whether classical, jazz, pop, or electronic—remains challenging. This adaptability requires a deep understanding of genre-specific characteristics, which are hard to encode in a single model.

Real-time music composition and interaction, such as improvisation in live performances or interactive music systems, present another set of challenges. These applications require the AI to

generate music on the fly, respond to human input, and adapt dynamically to changes in the musical environment. Achieving this level of interactivity and responsiveness still requires significant research and development. Moreover, collaboration between AI systems and human musicians introduces challenges in synchronizing AI-generated content with human performance. Ensuring AI compositions align with human musicians' timing, style, and emotional expression involves complex real-time processing and adaptability.

2.3.2. Specific Challenges with Transformer Models in Music

While Transformer and its variants have demonstrated excellence in automated authoring, they still have many challenges and issues to address. First, the Transformer model has high computational and memory requirements when dealing with long sequences, especially when generating long musical segments. It leads to longer training times and requires significant computational resources. While the relative attention mechanism partially alleviates this problem, further optimizations are needed to improve memory and computational efficiency.

Another challenge is that the Transformer model needs more precise control over musical style and emotion. While variants such as MuseNet and Jukebox can generate music in various styles, achieving precise control over specific styles or emotions remains challenging. In addition, generated music sometimes needs more creativity and novelty and is highly repetitive. Therefore, the model needs more innovative mechanisms to enhance the diversity and uniqueness of the music.

Future research can develop in several directions to address these issues. First, reducing computational complexity is crucial. Researchers should explore more efficient attention mechanisms, such as sparse or hierarchical attention, to reduce the computational complexity of processing long sequences. For example, fast autoregressive transformers with linear attention reduce the complexity from O (N^2) to O (N), significantly enhancing the model's performance on long sequences [10]. Moreover, Sparse Transformers utilize a decomposed attention model that significantly reduces memory and computational requirements while maintaining performance, enabling the model to process sequences of unprecedented length efficiently [11].

Further research into structural grokking could also provide insights into optimizing transformer depth and training duration better to capture hierarchical structures in musical compositions [12]. In addition, developing more efficient training algorithms and optimization techniques can improve models' training speed and generation efficiency. For example, Transformer-XL combines segment-level recursion and relative position encoding to handle more extended contexts with reduced computational complexity [9].

Second, it is crucial to enhance musical style and emotional expression. By combining emotion and style labelling, models train to generate music with specific emotions and styles, thus enriching the music's emotional richness and diversity. In addition, combining techniques such as Generative Adversarial Networks (GANs) or Variational Autoencoders (VAEs) can improve the models' ability to express emotions and finer details [6].

In addition, research on hybrid models has great potential. The hybrid model combines the advantages of RNNs and Transformers and can fully utilize both advantages. RNNs can deal with short-term and local dependencies, while Transformers can deal with long-term and global features. Exploring combinations of different types of neural networks can also help achieve higher-quality music generation.

Finally, applying Transformer models to practical scenarios is an essential direction for future research. For instance, using them for real-time music generation and interactive music systems can enable instant music creation during live performances and user interactions. Additionally, developing Transformer-based educational and creative tools can assist music educators and creators in conducting teaching and composition more efficiently.

By addressing both the general and Transformer-specific challenges, future research can pave the way for more sophisticated and emotionally resonant AI-generated music, ultimately enriching the field of automated music composition.

3. Conclusion

This research found that Transformer-based architectures (including Music Transformer, MuseNet, and Jukebox) significantly enhance the ability to generate high-quality, diverse, and stylistically rich musical compositions. These models effectively capture the complex temporal dependencies and hierarchical structures inherent in music beyond the limitations of traditional rule-based systems and RNNs. Furthermore, although researchers have yet to apply the Transformer-XL model in the domain of automatic music composition, it has the potential to generate coherent and complex long-form musical compositions. Its segment-level recursion mechanism and novel positional encoding scheme effectively capture long-term dependencies and address contextual fragmentation. Based on these characteristics, Transformer-XL promises to maintain the flow of information between musical segments, thereby generating scores with greater depth and coherence. Thus, Transformer's model can demonstrate potential for automated music composition tasks and provide better coherence and complexity to the music generated. Empirical results show that Transformer's model outperforms traditional RNN-based models regarding coherence, diversity, and ability to capture long-term dependencies. For example, Music Transformer has shown the ability to generate more pleasing and structurally coherent compositions than LSTM-based models. Thus, the success of these models in music generation opens up new avenues for research and applications, emphasizing the strong potential of Transformer in the creative field of music.

This study similarly demonstrates the effectiveness of Transformer variants in music generation. It shows that Transformer can be adapted and optimized for creative tasks beyond text and language processing. It emphasizes the potential of these models to revolutionize automated music creation, providing new tools and methods for musicians and technicians. It provides a comprehensive overview of the state of the art of Transformer modelling in music generation, thus encouraging further exploration and development of this promising field.

However, current research has limitations, including high computational and memory requirements and limited control over musical style and emotion. Future research should focus on optimizing computational efficiency, for example, exploring more efficient attention mechanisms such as sparse or layered attention. Enhancing the ability of models to express and control specific musical styles and emotions will also be critical. Future research should also consider developing hybrid models that integrate the strengths of RNNs and Transformers, thereby improving the quality and diversity of automatic music generation.

References

[1] Eck, D., & Schmidhuber, J. (2002). Finding temporal structure in music: Blues improvisation with LSTM recurrent networks. Proceedings of the 12th IEEE Workshop on Neural Networks for Signal Processing. DOI: 10.1109/NNSP.2002.1030094.

[2] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30. https://doi.org/10.48550/arXiv.1706.03762.

[3] Hadjeres, G., & Pachet, F. (2016). DeepBach: a steerable model for Bach chorales generation. https://doi.org/10.48550/arXiv.1612.01010.

[4] Huang, C. A., Vaswani, A., Uszkoreit, J., Shazeer, N., Simon, I., Hawthorne, C., & Eck, D. (2018). Music transformer. https://doi.org/10.48550/arXiv.1809.04281.

[5] Payne, Christine. "MuseNet." OpenAI, 25 Apr. 2019. Available at: openai.com/blog/musenet.

[6] Dhariwal, P., Jun, H., Payne, C., Kim, J. W., Radford, A., & Sutskever, I. (2020). Jukebox: A generative model for music. https://doi.org/10.48550/arXiv.2005.00341.

[7] Huang, C. A., Cooijmans, T., Roberts, A., Courville, A., & Eck, D. (2019). Counterpoint by convolution. https://doi.org/10.48550/arXiv.1903.07227.

[8] Shaw, P., Uszkoreit, J., & Vaswani, A. (2018). Self-Attention with Relative Position Representations. ArXiv: 1803.02155 [cs.CL]. https://doi.org/10.48550/arXiv.1803.02155.

[9] Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q., & Salakhutdinov, R. (2019). Transformer-XL: Attentive language models beyond a fixed-length context. https://doi.org/10.48550/arXiv.1901.02860.

[10] Katharopoulos, A., Vyas, A., Pappas, N., & Fleuret, F. (2020). Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention. ArXiv: 2006.16236 [cs.LG]. https://doi.org/10.48550/arXiv.2006.16236.

[11] Child, R., Gray, S., Radford, A., & Sutskever, I. (2019). Generating Long Sequences with Sparse Transformers. https://doi.org/10.48550/arXiv.1904.10509.

[12] Murty, S., Sharma, P., Andreas, J., & Manning, C. D. (2023). Grokking of Hierarchical Structure in Vanilla Transformers. ACL 2023. https://doi.org/10.48550/arXiv.2305.18741.