ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24APR1676

Generating Musical Sequences with Transformers

Nidhi Dewangan*
Computer Science Engineering
Bhilai Institute of Technology, Raipur
Raipur, India

Megha Singh
Computer Science Engineering
Bhilai Institute of Technology, Raipur
Raipur, India

Vijayant Verma
Computer Science Engineering
Bhilai Institute of Technology, Raipur
Raipur, India

Abstract:- **Transformers have significantly revolutionized the music-creation process by their ability to generate intricate and captivating musical arrangements. By analyzing patterns and connections within music data, transformers can produce new compositions with remarkable accuracy and originality. This study explores the internal mechanisms of transformers in music generation and highlights their potential for advancing the field of musical composition. The ability of transformers to capture extensive relationships and contextual information makes them highly suitable for tasks related to music generation. Through self-attention mechanisms, transformers effectively model the dependencies between different time intervals in a musical sequence, resulting in the production of coherent and melodious compositions. This paper delves into the specific architectural elements of transformers that enable them to comprehend and generate musical sequences while also exploring potential applications for transformer-based systems in various creative contexts - emphasizing on significant impact they could have on evolving techniques used during music composition.**

Keywords:- *Transformers, Music Generation, Compositions, Self-Attention Mechanism.*

I. INTRODUCTION

In the fast-evolving world of today, technology persists in transforming different sectors such as the art of music composition. With the emergence of Generative AI models such as transformers, the process of creating music has been deeply transformed. Gone is a day when composers rely solely on their creative abilities and musical knowledge to produce new melodies and harmonies. With the aid of deep learning and neural networks, composers can harness the power of machine-generated music to explore new horizons of creativity and innovation in their composition [9]. Through a comprehensive comprehension of diverse music genres and forms, the Generative AI model has the capability to produce distinctive alterations and amalgamations that have the potential to pose a challenge to music composers in forthcoming times. However, this injection of novelty and inspiration also serves as a catalyst for composers to further enhance their creativity and push the boundaries between their musical expressions. Transformers play a crucial role in this new wave of Generative AI, as they have shown immense potential in generating not only music but also images and text. Transformers, originally designed for language tasks, have been successfully applied to various generative tasks, including music generation. In recent times, the creation of music through the utilization of Transformers has captured a considerable amount of interest within the realm of research community.. The use of Transformers in music generation tasks is particularly intriguing due to their ability to model long-term dependencies and capture complex musical structures. The core architectural element of Transformers that enables them to comprehend and generate musical sequences is the self-attention mechanism. The self-attention mechanism allows Transformers to analyze the relationships between different elements in a sequence, whether it's words in a sentence or notes in a musical [1] composition.

Transformers, with their self-attention mechanisms, have proven to be highly effective in generating musical sequences. These models can capture intricate musical patterns, create diverse melodies, and compose complex harmonies. The flexibility and adaptability of Transformers allow them to generate music in various styles and genres, catering to the preferences of different composers and listeners. Transformers excel in analyzing the complex interplay between notes, chords, rhythms, and overall structure of music. This makes it possible to create musically coherent and stylistically consistent pieces [3]. In contrast to traditional methods, Transformers have proven to be effective in capturing long-term dependencies within a musical composition, which in turn enables the creation of music with a distinct sense of direction and avoids the appearance of repetitive patterns.

Transformer models, which are equipped with self-attention mechanisms, are leading the AI-driven revolution in melody generation. Their ability to identify intricate relationships within musical data enables them to create sequences with a remarkable level of coherence and creativity, a feat that was previously only accomplished by human composers [3] [2]. Auto regressive transformer models are praised for their ability to produce sequential notes that are aware of both proximate and extensive musical dependencies [3]. This compositional approach aligns with traditional human techniques, as it carefully synthesizes the intricate stylistic elements of the source material. By training on diverse datasets, such as the Lakh MIDI dataset, these AI models acquire the versatility needed to create a variety of musical styles, contributing to their remarkable adaptability and artistic range [6]. Transformers design enables the incorporation of distinct stylistic parameters, allowing the technology to generate music that corresponds to specific genres or imitates particular artists [2]. This tailored generative process harmonizes with the creative aspirations of composers, fostering a synergistic dynamic where human ingenuity informs and is amplified by machine intelligence.

Innovations in transformer-based music generation are exemplified by programs like OpenAI's Jukebox, which have revolutionized the landscape by generating fully-formed song structures, complete with nuanced lyrics, drawing from a vast array of genres and styles [2]. It is essential to acknowledge that the current trajectory of AI in music points towards a future where the relationship between technology and creativity becomes increasingly intertwined. Ongoing research in the field captivates curiosity and cumulatively advances the role of transformer models within the arts, as noted in studies like "A systematic review of artificial intelligence-based music generation" by Civit et al. and the in-depth exploration within "Music Transformer: Generating Music with Long-Term Structure" [3] [2].

II. LITERATURE REVIEW

The utilization of Transformer-like architectures in the training of music composition models has garnered considerable attention in recent times. These studies have significantly improved the Transformer's capacity in modeling music by implementing innovative techniques such as relative positional encoding schemes [10][8], cross-domain pre-training [7], and event token design [7][4].

In the current era, the advancement of intelligent music generation technology is progressing steadily, prompting researchers to explore the integration of neural networks with other models to facilitate the generation of multi-track music [11].

A significant development occurred in 2016 when Chu. H et al. unveiled an RNN-based hierarchical model (HRNN) that features a lower layer network responsible for generating melodies and an upper layer network dedicated to generating accompaniments for the melodies. Subsequently, in 2017, Yang. L et al. presented the Midi Net network [5], designed to produce complete melodies based on chord sequences or initial melodies, followed by further processing to generate multi-track music. Building upon these advancements, in 2017, Dong. H et al. presented the Muse GAN network [11], which combines convolutional neural networks and adversarial generation networks with the goal of generating multi-track music in the MIDI format.

Existing literature surrounding the assessment techniques for AI-generated music primarily classifies the evaluation methodologies into two main categories: subjective assessments and objective assessments. Subjective methods of evaluation involve the engagement of human participants to offer judgments based on specific criteria, including but not limited to musical quality, originality, and emotional resonance.

Many Researchers like Zhao et al. [2] have particularly highlighted the significance of utilizing listening experiments as a commonly adopted strategy in evaluating the harmonic compositions of AI-generated music. Due to the inherent subjectivity associated with the perception of music, establishing precise standardization in these assessments poses a formidable challenge, as articulated by Yamshchikov et al. [5]. However, advancements in machine learning and artificial intelligence offer promising avenues for automating certain aspects of music evaluation, potentially mitigating the challenges of subjective assessment and enhancing the efficiency of the evaluation process.

III. METHODOLOGY

The subsequent section describes the methodological approach employed to examine the capabilities of transformers in music production. The flow diagram in Fig.1 illustrates the process operation. The techniques for data acquisition and preprocessing used in preparing the music dataset for model training will be detailed. Besides, the specific transformer architecture chosen, along with its hyperparameter settings and training methodology, will be addressed. Lastly, an analysis will be conducted to evaluate the quality of the produced musical sequences.
`

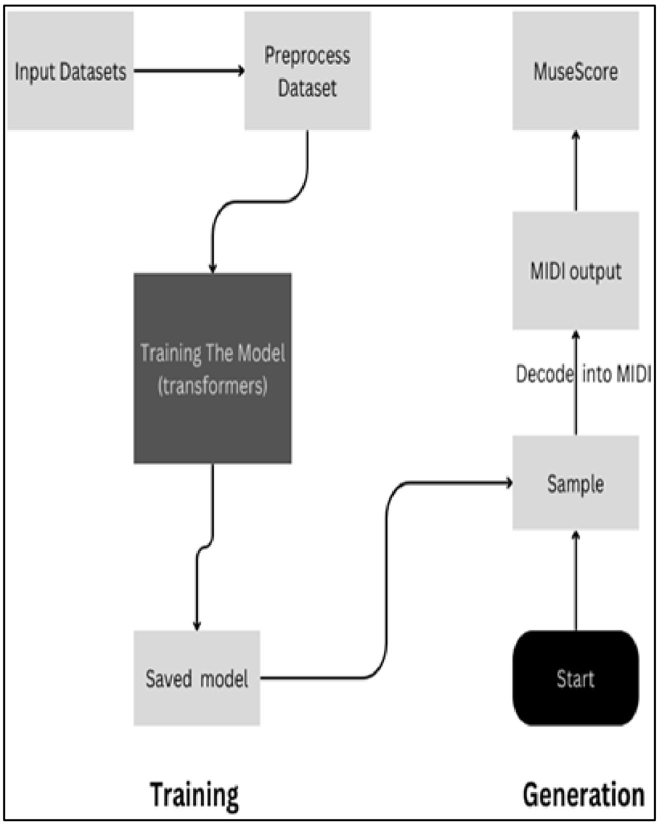


Fig. 1. Workflow Diagram of the Musical Sequence Generation

A. Data Collection And Preprocessing

- *Data Collection*

The musical dataset utilized in this study was sourced from the publicly accessible MAESTRO dataset. This dataset, which was curated by Google AI, is composed of piano performances that have been meticulously aligned with corresponding audio recordings. With a collection of over

200 hours of piano performances recorded using Yamaha Disklavier pianos, this dataset includes detailed MIDI data that captures keystroke force and pedal usage. The precise synchronization of audio and note data is crucial for the effective training of machine learning models in music generation. The MAESTRO dataset covers a wide array of musical styles and genres, thanks to its collaboration with the International Piano-e-Competition. This diverse range enables the model to learn from a broad spectrum of musical elements such as structures, harmonies, and rhythms, thus improving its capacity to create a variety of musical pieces. Each piece in the dataset is carefully segmented and annotated with information on the composer, title, and performance year. This systematic arrangement aids in the efficient retrieval and exploration of specific musical segments for various training and testing purposes. The splitting of the MAESTRO dataset into training, validation, and test sets is a typical strategy in machine learning to prevent overfitting and enhance the model's ability to generalize.

➢ *Data Preprocessing*

Upon acquisition of the extensive musical data from the MAESTRO dataset, preprocessing procedures were initiated to ready it for the training of the transformer model. The primary objective of these procedures was to convert the raw musical data into a form that is appropriate for the architecture and learning mechanisms of the model.

- MIDI Representation: In relation to MIDI Representation, the compositions found in the MAESTRO dataset are presented in MIDI form . MIDI data provides a symbolic depiction of music, encompassing details regarding notes, their pitch, duration, and velocity (the intensity of keystrokes). This symbolic structure was preserved for our objectives due to its capacity to enable the model to concentrate on the fundamental musical framework and connections between notes.

- Pipeline Preprocessing : The data preprocessing pipeline is vital for preparing musical data in the form of MIDI files to train a music generation model. It begins by parsing MIDI files and converting them into a format that is suitable for further processing. Next, the pipeline extracts detailed note information such as pitch, start/end times, and velocity, while also considering sustain pedal activity for accurate timing. The extracted notes are then divided into training and validation sets, with the option to apply data augmentation techniques like stretching note durations and transposing pitches to create a more diverse training dataset. Finally, longer sequences are cut into smaller chunks of a fixed length, and both timing and velocity information are converted into discrete values. This process ensures consistent data representation, prepares it for the model's format, and offers flexibility for data augmentation, ultimately transforming raw MIDI files into a structured format that enables the music generation model to learn and create new pieces.

- Sequence Encoder: The transformer model and preprocessed music are connected by the Sequence Encoder class. It converts previously processed musical note sequences into a format that the model can comprehend. Creating an event language to express note onsets, offsets, time shifts for progressive progression, and even velocity changes for dynamics are all part of this metamorphosis. Next, the encoder gives each type of event and its matching value (such as pitch for note onsets) a distinct numerical number. The model can effectively handle the intricate temporal and dynamic interactions seen in the music thanks to its numerical encoding. The encoder records the notes in the piece's order as well as its dynamic changes by looping over each sequence and turning it into a list of these event codes. Additionally, the encoder provides flexibility by trimming excessively long sequences or rejecting extremely short ones in order to comply with the input requirements of the model. In the end, this procedure sets up the information such that the transformer model may discover the underlying musical patterns and create new compositions that incorporate the discovered stylistic features.

-

B. *Transformer Model Architecture*

The core architecture of the music generation model is defined in the Music Transformer class. It's a generative and autoregressive transformer that predicts the next element in a musical sequence one step at a time, depending on the previous elements. Here's a breakdown of the key components.

- Initialization: Takes various input hyperparameters (number of unique tokens, sequence length, embedding dimensionality, etc.) as input. Creates an embedding layer that maps each token in the vocabulary to a dense vector representation. Optionally implements positional encoding.

- Forward Pass: The input sequence of tokens and an optional mask are taken as input. The input sequence is embedded using the embedding layer. Positional encoding (if enabled) is used to inject information about the order of tokens within the sequence. The encoded sequence passes through each decoder layer in the stack. To capture long-range dependencies and extract meaningful features from the sequence, the decoder layers perform a combination of self-attention and feed-forward operations. Returns the output logits representing the probabilities of the next token in the sequence.

- Attention: The Multiheaded Attention class implements self-attention as a core component within the decoder layer. Here's a simplified explanation of its functionality

- Split Inputs and Linear Layers: The input sequence is divided into separate queries, keys, and values for each of the multiple attention heads.

Linear layers were applied to transform these elements into lower dimensions suitable for attention calculation.

- Relative Positional Embeddings : If enabled, relative positional embeddings are considered to capture the distance between elements in the sequence. This inputs information about the order beyond the absolute position of each token.

- Scaled Dot-Product Attention : Attention scores are calculated by computing the dot product between queries and keys, scaled by a constant factor. Higher scores indicate a stronger link between the query and the corresponding key.

- Masking: A mask is then applied to prevent the model from attending to future tokens in the sequence, while maintaining the autoregressive nature of generation. An additional mask can be provided that explicitly excludes specific elements from attention calculations.

- Soft max and dropout are: Attention scores are converted to probabilities using soft max , which indicates the weight or importance of each value in the sequence. Dropout was applied for regularization to improve model stability and prevent overfitting.

- Weighted Average and Recombination: The attention probabilities are used to compute a weighted average of the values, which effectively corresponds to the most relevant parts of the sequence based on calculated scores .To produce the final output, the results from all attention heads are combined using a linear layer.

C. *Traning Procedure*

Using an iterative process (epochs), the Music Transformer model is trained on the prepared training data. Within each epoch: Mini-batches of training data are separated according to batch_size. The input music sequence, or seed, is fed into the model for every batch. The probability of the subsequent element in the sequence are predicted by the model. A loss function (e.g., cross-entropy with label smoothing) is used to compare the projected probability with the actual next element. To minimize the loss, the model's weights are updated by back propagating the loss via the network. To track the model's capacity for generalization, validation loss is computed on a regular basis on the validation set. Until a predetermined number of epochs (n_epochs) is reached or a suitable validation loss is attained, the training process is continued.

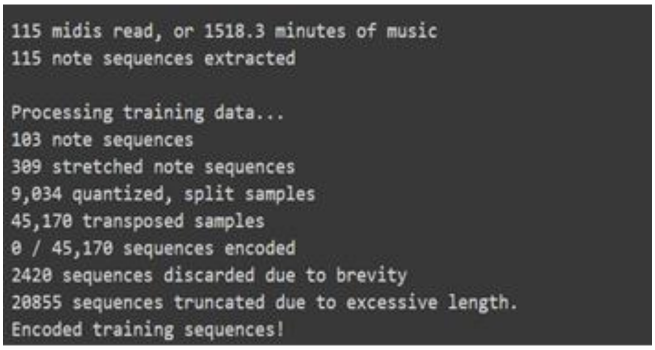
D. *Music Sequence Generation*

Using a seed, which is a brief musical excerpt, the model may be trained to create new musical sequences. The seed sequence is input into the model during generation. Using the patterns it has learnt from the training set of data, the model predicts the subsequent element in the sequence. The resulting sequence has the expected element attached to it. After carrying out this procedure repeatedly for a predetermined amount of time (sample_ length), a freshly composed musical composition is produced.

Temperature parameter Regulates the produced sequence's degree of unpredictability. Higher temperatures provide a wider range of results, some of which may be unexpected.

IV. Result And Discussion

Training was conducted using the Google Colab platform, where the machine learning models underwent training sessions lasting up to 100 epochs, or until reaching a point where there was no further improvement observed in the loss function over a span exceeding 5 epochs. Following the completion of the training process for the generative model Fig.2, a series of musical compositions were produced. To provide a representation of the outcomes, a selection of music pieces was randomly chosen from the generated results and subjected to visualization techniques. Illustrated in the accompanying Fig.3 ,the musical sequences generated exhibited a unique quality and were subsequently executed in the software application known as Musescore. Noteworthy is the fact that the duration of each training epoch reached its peak at 12 minutes, specifically in the case of the most intricate version of the initial architectural design utilized. Furthermore, the analysis of the outcomes revealed that the maximum level of validation accuracy achieved was recorded at 0.1, accompanied by a validation loss metric of 3.20.



```
115 midis read, or 1518.3 minutes of music
115 note sequences extracted

Processing training data...
103 note sequences
309 stretched note sequences
9,034 quantized, split samples
45,170 transposed samples
0 / 45,170 sequences encoded
2420 sequences discarded due to brevity
20855 sequences truncated due to excessive length.
Encoded training sequences!
```

Fig. 2. Parsing Training dataset

Training an epoch lasted a duration of 150 minutes, a significant amount of time required for the model to undergo the necessary learning process.

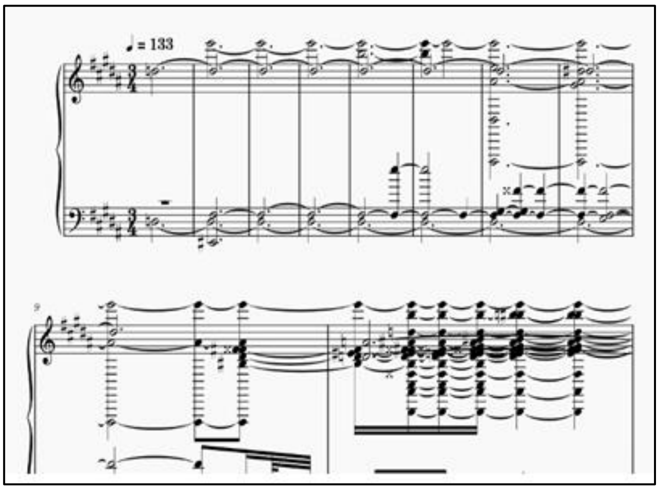


Fig. 3. Generated sample sequences

A period that could extend over several days until the model reaches its full potential through thorough training.

This particular duration of time proves to be highly impractical when considering the training of models on the entirety of the dataset, presenting a challenge in terms of efficiency and resource allocation.

V. CONCLUSION AND FUTURE WORK

The paper introduces the Transformer model for enhancing machine-generated music quality, along with metrics highlighting its shortcomings. These metrics reveal issues like erratic pitch usage, inconsistent grooving patterns, and lack of repetitive structures, indicating the Transformer's limitations in composing music. Music generation with AI faces challenges like data availability, computational resources, and evaluation metrics, but tools like PrettyMIDI and PyTorch can help overcome these obstacles. The Transformer model shows promise in generating music across various styles and genres, offering new creative possibilities for musicians and composers. AI music generation has the potential to revolutionize the industry, with opportunities for innovation and experimentation to drive future developments.

REFERENCES

[1]. Douzon, T., Duffner, S., García, C., & Espinas, J. (2023, January 1). *Long-Range Transformer Architectures for Document Understanding*. Lecture Notes in Computer Science. https://doi.org/10.1007/978-3-031-41501-2_4

[2]. Wang, L., Zhao, Z., Liu, H., Pang, J., Qin, Y., & Wu, Q. (2022, November 16). *A Review of Intelligent Music Generation Systems*. arXiv.org. https://arxiv.org/abs/2211.09124

[3]. Muhamed, A., Li, L., Shi, X., Yaddanapudi, S., Chi, W., Jackson, D., Suresh, R., Lipton, Z. C., & Smola, A. J. (2021, May 18). *Symbolic Music Generation with Transformer-GANs*. Proceedings of the . . . AAAI Conference on Artificial Intelligence. https://doi.org/10.1609/aaai.v35i1.16117

[4]. Ji, S., Luo, J., & Yang, X. (2020, November 13). *A Comprehensive Survey on Deep Music Generation: Multi-level Representations, Algorithms, Evaluations, and Future Directions*. arXiv.org. https://arxiv.org/abs/2011.06801

[5]. Yamshchikov, I. P., & Tikhonov, A. (2020, November 2). *Music generation with variational recurrent autoencoder supported by history*. SN Applied Sciences/SN Applied Sciences. https://doi.org/10.1007/s42452-020-03715-w

[6]. Wu, S. L., & Yang, Y. H. (2020, August 4). *The Jazz Transformer on the Front Line: Exploring the Shortcomings of AI-composed Music through Quantitative Measures*. arXiv.org. https://arxiv.org/abs/2008.01307

[7]. Huang, Y. S., & Yang, Y. H. (2020, February 1). *Pop Music Transformer: Beat-based Modeling and Generation of Expressive Pop Piano Compositions*. arXiv.org. https://arxiv.org/abs/2002.00212

[8]. Donahue, C., Mao, H. H., Li, Y. E., Cottrell, G. W., & McAuley, J. (2019, July 10). *LakhNES: Improving multi-instrumental music generation with cross-domain pre-training*. arXiv.org. https://arxiv.org/abs/1907.04868

[9]. Briot, J. P., & Pachet, F. (2018, October 16). *Deep learning for music generation: challenges and directions*. Neural Computing & Applications. https://doi.org/10.1007/s00521-018-3813-6

[10]. Huang, C. Z. A., Vaswani, A., Uszkoreit, J., Shazeer, N., Simon, I., Hawthorne, C., Dai, A. M., Hoffman, M. D., Dinculescu, M., & Eck, D. (2018, September 12). *Music Transformer*. arXiv.org. https://arxiv.org/abs/1809.04281

[11]. Dong, H. W., Hsiao, W. Y., Yang, L. C., & Yang, Y. H. (2017, September 19). *MuseGAN: Multi-track Sequential Generative Adversarial Networks for Symbolic Music Generation and Accompaniment*. arXiv.org. https://arxiv.org/abs/1709.06298

[12]. Yang, L. C., Chou, S. Y., & Yang, Y. H. (2017, March 31). *MidiNet: A Convolutional Generative Adversarial Network for Symbolic-domain Music Generation*. arXiv.org. https://arxiv.org/abs/1703.10847