International Journal on Smart & Sustainable Intelligent Computing
Vol. 01, Iss. 02, S. No. 005, pp. 58-69, October 2024
ISSN (E):3048-8508

Applying Recurrent Neural Networks with integrated Attention Mechanism and Transformer Model for Automated Music Generation

Aditya Kumar, Ankita Lal
Kalinga Institute of Industrial Technology, India
aditya007kkr@gmail.com, ankitalal207@gmail.com

How to cite this paper : Aditya Kumar, Ankita Lal "Applying Recurrent Neural Networks with integrated Attention Mechanism and Transformer Model for Automated Music Generation," *International Journal on Smart & Sustainable Intelligent Computing*, Vol. no. 01, Iss. No 02, pp. 58-69, October 2024.

Received: 25/08/2024
Revised: 30/09/2024
Accepted: 20/10/2024
Published: 31/10/2024

Abstract

One of the key applications of artificial intelligence has come out in the form of automated music generation, that is a hybrid of creativity and computational models. The main purpose of this research is to develop and understand the blend of advanced Recurrent Neural Networks(RNNs) architecture with attention mechanisms to improve the limitations of ongoing RNNs in capitalizing long term dependencies. The main targets of the model are on relevant segments of input sequences, ensuring enhanced consistency and structural strength in the generated music industry. By comparison of enhanced RNN with a Transformer based model known for its exceptional capacity to model long-range dependencies through self-assessment. Additionally to filter rhythmic exactness and style, a beat-level segmentation method is implemented into this process. The structural composition of the generated outputs is examined using a Self-Similarity Matrix (SSM), which balances between reiteration and diversity

Keywords

Music Generation, Recurrent Neural Network (RNN), Attention Mechanism, Transformer Model, Self-Similarity Matrix (SSM), Beat Level Segmentation

1. Introduction

Deep learning has a remarkable impact in our everyday life, from chatbots to voice assistants. Learning from given data sets, analyzing the pattern and being able to create a new one is just magical and a topic of research. Deep learning has revolutionized and automated various industries including arts. Producing music is not an easy task, it takes lots of time and effort, it not only helps musicians but also people with no knowledge about music can also produce it using RNN. A “Recurrent Neural Network(RNN)” is a deep learning model for processing data in sequential order. It processes data step by step, output from previous steps are fed back, so that it can learn temporal patterns. One common drawback of RNN is that it may suffer from vanishing or exploding gradients. This limitation makes the model inefficient to learn long-term dependencies.

Attention mechanisms are added to RNN and its variants like Long Short Term Memory (LSTM) and Gated Recurrent Units (GRU) to avoid these limitations. It has selective focus, to focus on a significant part of the sequence and that thereby captures nitty gritty patterns of the composition. Musical Instrument Digital Interface (MIDI) are the input files, they are like the building block for music with detailed notes and timing. In contrast to RNN, a transformer based model uses a self attention mechanism. Instead of processing data step by step, they capture everything at once, which makes the model effective to recognize complex relationships. While RNN is better for capturing both long term and short term patterns when integrated with attention mechanisms, transformer based models are good at handling complex compositions.

The main contribution of the work are as follows:

- Our model improved the Recurrent Neural Network (RNN) by integrating attention mechanisms. It removes the limitations such as vanishing or exploding gradients and learning long term dependencies. This helps in more accurate music generation along with focusing more on the relevant part of the sequence.

- Our paper compares both the leading models for music generation which are RNN and transformers. It includes the structural part and the basic working of the models, highlighting the pros and cons of both the approaches.

- It includes beat level segmentation which basically means breaking music into small chunks to analyze it more efficiently. Extracting data from pre trained data helps to capture important features. This makes the generated music have a more natural style and beats.

- To make the music more enjoyable, we implemented the Self Similarity Matrix, it basically compares different parts of the music to see the similarity and difference between them. It then uses the perfect combination of recurrence and alteration to make the music more interesting and not boring.

2. RELTAED WORKS

Conventional automatic music generation systems often focus on single-track generation, which lacks the capability to address harmony, rhythm alignment, and overall structure. This limitation underscores the need for advanced methods to enable more sophisticated and compatible multi-track music compositions [1]. Recurrent Neural Networks (RNNs) have been highlighted for their importance in processing sequential data, transforming machine learning through effective analysis of texts and speeches. Advancements such as LSTM, GRU, BiLSTM, Stacked LSTM, and Peephole LSTM have been detailed. LSTM excels at handling long-term dependencies, GRU simplifies the process, BiLSTM processes bidirectionally, Stacked LSTM employs layered processing for enhanced performance, and Peephole LSTM improves accuracy through additional connections. These technologies have been applied in time-series forecasting, decision-making, anomaly detection, and speech recognition [2]. RNNs have been examined for their roles in various applications, assessing their effectiveness and the challenges encountered in practical implementations [3].

While music is a universal source of enjoyment, current music creation tools often cater to experienced musicians. A novel RNN-based approach has been proposed to empower novices by training on existing melodies or instrumentals, making music creation accessible to amateurs and professionals without requiring physical instrumentation skills [4]. Recent studies have developed automatic music generators utilizing MIDI files as input. Research employing LSTM and GRU networks demonstrated that a double-stacked GRU model outperformed others in mimicking composer patterns, achieving a 70% recall score through subjective evaluation [5]. Personalized assessments of generated music revealed its listenability and appeal, with the double-stacked GRU layer achieving a rating of 6.85 out of 10. Deep learning models have also been shown to improve music spectrogram structure, enabling better classification of music artists while exploring factors like music length and production styles [6].

Computer-generated music has vast applications, often relying on AI and ML for producing matching music scores. Few music synthesis models utilize artificial neural networks (ANNs), some struggling with transposition invariance. To address this, RNN-based melody synthesis technology has been developed, which extracts acoustic features and adopts sequence-sequence models for melody and singing synthesis. Experimental results validate the model's effectiveness [7]. The rise of online video-sharing platforms has increased the demand for well-matched background music in filmmaking. Effective storytelling relies heavily on background music to enhance the narrative. Recent studies on music generation have focused on creating melodies, covering pitch, rhythm, duration, and pauses [8].

Transformer-based models, such as MTMG, have been proposed for generating multi-track music, including piano, guitar, and drum tracks. These models leverage transformer architecture with innovations like pairwise learning and GPT-based sequence prediction, demonstrating superior musical representation in evaluations [9]. Transformers are further advancing music generation, particularly in classical and pop compositions. For instance, the "Pop Music Transformer" represents inputs with clear beat-bar phrase structures, achieving improved beat alignment and musical structure compared to earlier models [10].

3. PROPOSED METHODOLOGY

The main motto of our research activity was to propose an activated Recurrent Neural Network (RNN) architecture boosted with attention mechanisms and inspect its performance with Transformer-based models for automated music generation. The major aspects of the proposed model include:

- The conventional RNN design is elevated by incorporating attentional capabilities, enabling selective concentration on pertinent segments of musical sequences. This refinement mitigates the vanishing gradient problem and enhances the model's capability to apprehend distant relationships, deriving in more coherent and structurally accurate music generations.
- Transformer-based Model: A Transformer model with self-attention is introduced as a serpoint for comparison. The Transformer model advances in decoding long-term dependencies, enabling creation of more varied, intricate, and statistically consistent composition.
- A novel beat-level segmentation technique boosts the rhythmic accuracy and stylistic authenticity of generated music. By dividing musical sequences into discrete, beat-aligned segments, the model gains finer control over genre-specific and period-dependent rhythmic patterns, yielding compositions that more closely emulate the rhythmic nuances of diverse musical styles.

- Self-Similarity Matrix (SSM) Evaluation: The use of the Self-Similarity Matrix (SSM) is to evaluate the structural quality of the generated music. This matrix computes the balance between repetition and variation within the generated music. It also ensures that the compositions are both engaging and musically coherent.

The overall architecture as shown in figure 1 involves the following flow:

- Input: The input to the model is a sequence of musical compositions in the form of MIDI files. These sequences are then preprocessed into embeddings, which are fed into the model.
- Music Generation Process: The enhanced RNN or Transformer model processes the input sequence. For the RNN, attention mechanisms are applied to focus on important parts of the input sequence, while the Transformer model uses self-attention layers to capture long-range dependencies.
- Beat-Level Segmentation: The music generated by both models is analysed and segmented into beat-level representations in order to improve rhythmic and stylistic accuracy.
- Output: The generated music is exported as MIDI files or audio tracks that can be further analysed using the Self-Similarity Matrix (SSM) for quality assessment.
- Evaluation: The SSM is used to elaborate the rational structure of the generated music by analysing its balance of variation and repetition.

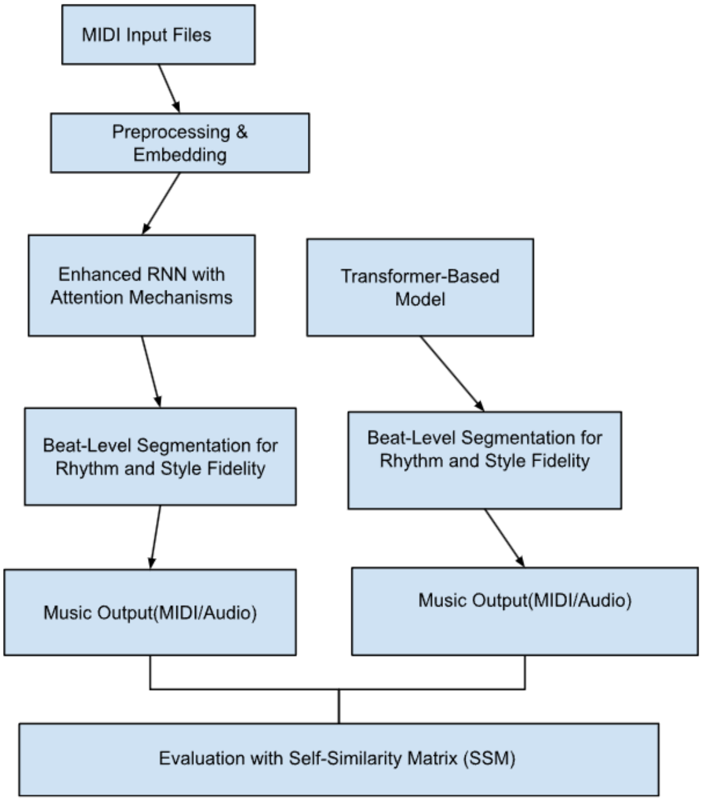
MIDI Input Files
Preprocessing & Embedding
Enhanced RNN with Attention Mechanisms
Transformer-Based Model
Beat-Level Segmentation for Rhythm and Style Fidelity
Beat-Level Segmentation for Rhythm and Style Fidelity
Music Output(MIDI/Audio)
Music Output(MIDI/Audio)
Evaluation with Self-Similarity Matrix (SSM)

Figure 1. Proposed model for music generation

Traditional RNNs suffer from the problem of vanishing gradients, making them less effective for sequences with long-term dependencies. To address this, the attention mechanism was incorporated. Transformers utilize self-attention mechanisms to model global dependencies in sequences, avoiding the sequential bottleneck of RNNs.

1. Self-Attention

Self-attention focus on a significant part of the sequence with respect to one another:

Attention(Q,K,V)=softmax(QK^T/dk^(½))V

where, Q=Query, part of sequence asking for information

K= Key, part being compared to query

V= Value, actual value one needs to focus on.

2. **Multi-Head Attention**

Multihead attention improves the ability of the model to focus on different parts of the sequence simultaneously:

MultiHead(Q,K,V)=Concat(head-1,…,head-h)Wo

where, Heads= Multiple independent attention mechanisms, learn to focus on different parts of the input.

Cocat= Output of all the heads are joined together

W0= Weight matrix to make the result into single output

Computing Each Head:

head-i=Attention(QW^Q-i,KW^K-i,VW^V-i)

where W^Q-i, W^K-i, W^V-i= Different weight matrix for head-i.

3. **Position Encoding**

Transformers use positional encodings to inject sequential information into the input:

For even dimensions (2i):

PE(pos,2i)=sin(pos/10000^(2i/d))

For odd dimensions (2i+1):

PE(pos,2i+1)=cos(pos/10000^(2i/d))

4. **Feedforward Network**

After the attention mechanism, a feedforward network is applied:

FFN(x)=ReLU(xW1+b1)W2+b2

where, x= output from attention layer and input for feedforward network.

W1, W2= Weight matrices

b1, b2= Learnable bias terms for each transformation.

ReLU= Non linear activation function.

.

5. Training Objective

The loss function for both models is based on cross-entropy:

L=−(t=1 to T∑logP(yt|x1:T))

where, T: Total number of time steps

yt: The correct output at time step t.

P(yt|x 1:T): The model's predicted probability for the correct output yt given the input sequence x 1:T.

The evaluation metrics used in the process include:

- **Self-Similarity Matrix (SSM):**

  To evaluate structural coherence, the SSM is computed as:

  S(i,j)=cos⁡(Embedding(i),Embedding(j))S(i, j) = \cos(\text{Embedding}(i), \text{Embedding}(j))S(i,j)=cos(Embedding(i),Embedding(j))

  - S(i,j)S(i, j)S(i,j): Similarity between segments iii and jjj.
  - Embedding(i)\text{Embedding}(i)Embedding(i): Vector representation of the iii-th segment.

- **Melodic Coherence:**

  Measured using average pitch deviation across generated sequences.

- **Rhythmic Complexity:**

  Quantified using syncopation and beat alignment metrics.

4. RESULTS AND ANALYSIS

The evaluation of the enhanced RNN model with attention mechanisms and Transformer-based models was conducted using a large dataset of MIDI files encompassing diverse musical styles. The results were analyzed based on melodic coherence, rhythmic complexity, and structural progression, with key findings. The incorporation of attention mechanisms significantly improved the RNN model's capability to handle long term dependencies, deriving in smoother transitions and more coherent compositions compared to traditional RNNs. Beat level segmentation improved the model's capability to produce more humanly and enjoyable music as it enhanced its rhythmic alignment by breaking the music into smaller units. This technique is best suited to produce a particular genre.

Transformer created more structured and complex music. It is faster and handles larger datasets better. Both models balanced repetition and alternation to make it more enjoyable and interesting. It can mimic genres more accurately. The analysis is shown in figure 2.

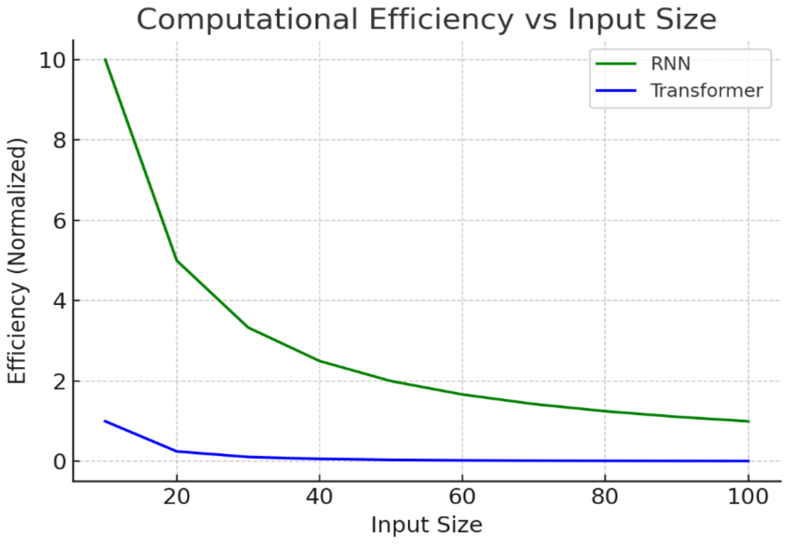

Figure 2. Efficiency vs Input Size

RNNs are more efficient than transformers when it comes to handling small input sizes because of their sequential processing which requires less memory. On the other hand, Transformers have quadratic complexities, making them less effective for short sequences even though they excel at handling long term dependencies.

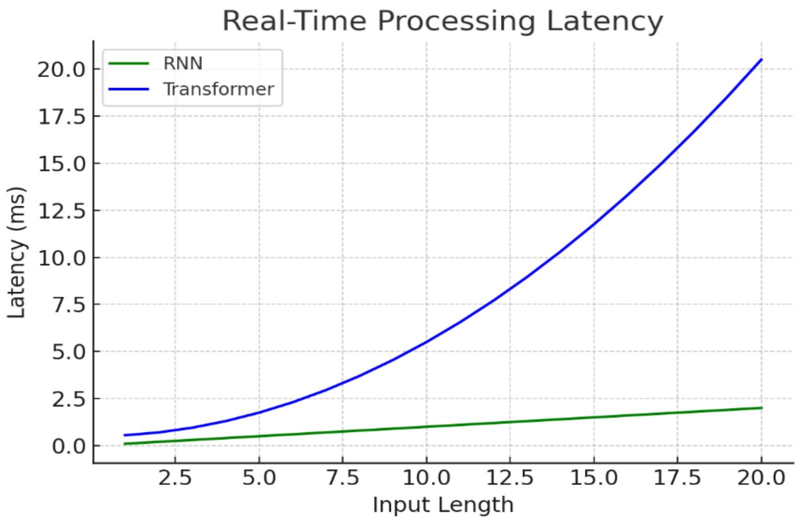

Figure 3. Latency vs Input Length

RNNs process data in a step by step order, which reduces latency for real time tasks, whereas transformers require complete input at once, leading to higher latency as input size increases. The overall analysis is depicted in figure 3.

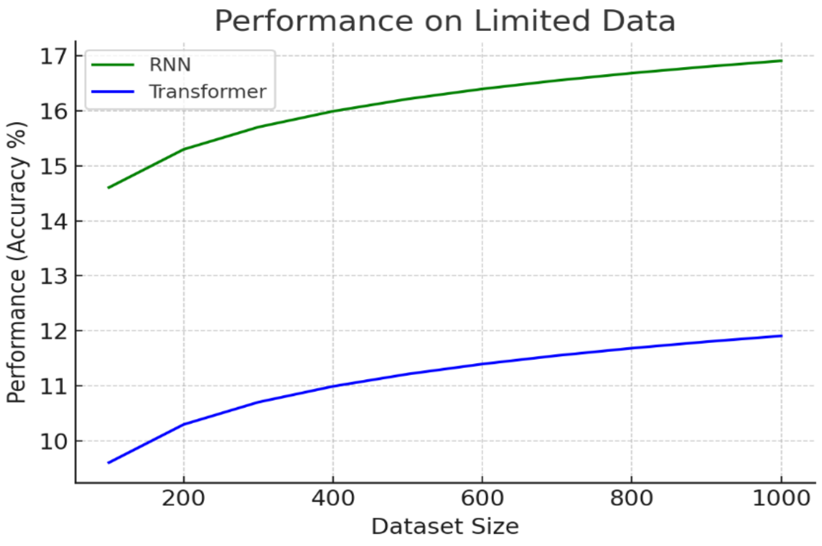

Figure 4. Performance vs   Dataset Size

RNNs have simple architecture and limited training parameters. Therefore, it performs better on smaller datasets. Contrast to this, transformers need more data to recognise patterns and work effectively, limited data make them underperform. This outcome is shown in figure 4.

5. CONCLUSION

This research dives deep into the new advancements in the music generation field by comparing Recurrent Neural Networks (RNNs) augmented with attention mechanisms and Transformer-based models. One can generate context-driven and logical results by the use of attention mechanisms with a mix of RNNs, which is broadly described by this study. But if we talk about long range dependencies and producing thematically rich compositions along with structural coherence support, then surely the Transformer based models have an upper hand on RNNs. The improvement in beat-level segmentation has increased the sleekness and rhythmic attractiveness. To assess structural strength, by providing a robust mechanism, the Self-Similarity Matrix (SSM) evaluation metric has been developed. These developments have implemented some practical gadgets and ideologues for enhancing the domain of automated music generation. Talking about the creative domains, these findings have opened new ways for the transformative potential of deep learning and created a hope for future experiments and expansions in the generation of music. With the attention mechanism the RNNs facilitate a hard foundation stone for incrementing orthodox processes. The Transformer-based models showcase the state-of-the-art, which is pushing the limits of what AI can do in the realm of arts and entertainment. A blind of hybrid model and the power of RNNs and Transformers can bring wonders in the upcoming generation as well as optimise new learning techniques and genre-specific customisation to increase the diversity and quality of Ai-Generated music.

References

[1] Jiang, R., & Mou, X. (2024), "The Analysis of Multi-Track Music Generation With Deep Learning Models in Music Production Process," IEEE Access, 12, 110322-110330.

[2] Mienye, Ibomoiye Domor, Theo G. Swart, and George Obaido. "Recurrent neural networks: A comprehensive review of architectures, variants, and applications." 2 Inf., 15, 517.

[3] Ilhan, F., Karaahmetoglu, O., Balaban, I., & Kozat, S.S. (2020). "Markovian RNN: An Adaptive Time Series Prediction Network With HMM-Based Switching for Nonstationary Environments. IEEE Transactions on Neural Networks and Learning Systems," 34, 715-728.

[4] Sajad, S., Dharshika, S., & Meleet, M. (2021). “Music Generation for Novices Using Recurrent Neural Network (RNN). 2021 International Conference on Innovative Computing,” Intelligent Communication and Smart Electrical Systems (ICSES), 1-6.

[5] Gunawan, A.A., Iman, A.P., & Suhartono, D. (2020). “Automatic Music Generator Using Recurrent Neural Network.” Int. J. Comput. Intell. Syst., 13, 645-654.

[6] Nasrullah, Z., & Zhao, Y. (2019). “Music Artist Classification with Convolutional Recurrent Neural Networks.” 2019 International Joint Conference on Neural Networks (IJCNN), 1-8.

[7] Zhang, Y., & Li, Z. (2021). “Automatic Synthesis Technology of Music Teaching Melodies Based on Recurrent Neural Network.” Sci. Program., 2021, 1704995:1-1704995:10.

[8] Hsu, J., & Chang, S. (2021). “Generating Music Transition by Using a Transformer-Based Model. Electronics.”

[9] Jin, C., Wang, T., Liu, S., Yun, T., Li, J., Li, X., & Lui, S. (2020). “A Transformer-Based Model for Multi-Track Music Generation.” Int. J. Multim. Data Eng. Manag., 11, 36-54.

[10] Huang, Y., & Yang, Y. (2020). Pop Music Transformer: Beat-based Modeling and Generation of Expressive Pop Piano Compositions. Proceedings of the 28th ACM International Conference on Multimedia.