Research on music generation based on Transformer

Xuemei Bai[1st]
School of Electronic Information Engineering,
Changchun University of Science and Technology,
Changchun 130012, China
e-mail: baixm@cust.edu.cn

Huiyuan Zhao[2nd]
School of Electronic Information Engineering,
Changchun University of Science and Technology,
Changchun 130012, China
e-mail: 2022100706@mails.cust.edu.cn

Chenjie Zhang[3rd, *]
School of Electronic Information Engineering,
Changchun University of Science and Technology,
Changchun 130012, China
*Corresponding author e-mail: zhangcj@cust.edu.cn

Hanping Hu[4th]
School of Computer Science and Technology,
Changchun University of Science and Technology,
Changchun 130012, China
e-mail: hhp@cust.edu.cn

Abstract **Music generation refers to the use of computers to create music through certain algorithms or processes with minimal human intervention, and is an important research direction in the field of artificial intelligence. However, music generation is an unsolved problem facing many challenges, and how to generate high-quality music samples has always been a research hotspot, for this reason, this paper proposes a music generation model based on improved Transformer. Specifically, this paper firstly preprocesses the music data and converts the music data into sequences for the model to process and learn. Secondly, a gated recurrent unit (GRU) is introduced to improve the Transformer, and the feature extraction capability of the Transformer to capture global dependencies and the time series modelling capability of the GRU are used to better understand the contextual relationships of the sequence data, and to generate high-quality music that is closer to real music. Finally, experimental verification of the music generated by the proposed method is carried out in this paper, and the results prove the effectiveness of the method.**

Keywords music generation; automatic composition; deep learning; Transformer; attention mechanism

I. INTRODUCTION

Music has always been an indispensable part of human culture, which is not only a form of entertainment, but also carries the inheritance of emotion, thought and culture. With the rapid development of computer science and artificial intelligence technology, music generation technology has also received extensive attention and research.

Music generation, also known as algorithmic composition, usually refers to the use of computers to create music through certain algorithms or processes with minimal human intervention [1]. Traditional music composition requires creators to have deep musical literacy and creative experience, as well as an understanding of music theory and structure, and at the same time, music composition is very time-consuming and labour-intensive, while algorithmic composition using deep learning combines human creativity, emotional expression and computer computational capabilities and other technologies, breaking through the constraints of the expertise of human composition to create newer and more exotic musical effects.

Early researchers utilised Recurrent Neural Networks (RNN) for music generation, the Magenta team at Google Brain proposed the Melody-RNN model [2] to improve the ability of RNNs to learn long term structures, and Tiwari [3] utilised Long Short-Term Memory (LSTM) networks in order to produce coherent and pleasing melodies similar to the original data distribution. As deep learning techniques continue to evolve, powerful deep generative models such as GANs and VAEs are emerging. Liu [4] reviewed the application of algorithm-based approaches to intelligent music composition and introduced the role of Generative Adversarial Networks (GANs) for music generation. Sharma and Bvuma [5] explored the potential of GANs for creative applications, emphasising their role in music generation. Transformers constitute a relatively new architecture [6] and the model has also shown its great potential in music generation. Huang et al. successfully applied Transformer for the first time to create music with long-term structure [7]. In [8], a model called Pop Music Transformer was proposed to generate popular piano music. The model uses a beat-based representation of music. The generated tracks are evaluated by an expert and a general listener and both groups prefer alternative architectures. Zhang [9] proposed a template-based learning adversarial Transformer for symbolic music generation that improves the quality and diversity of music generation.

Relevant studies have shown that the application of deep learning in music creation is feasible and effective, but improving the music generation effect is still the key. To this end, this paper proposes a music generation model based on Transformer. The model introduces a GRU network in the encoder of Transformer, so that while the attention mechanism focuses on global information, the GRU effectively captures local dependencies, enhances the modelling ability of temporal dependencies through the self-attention mechanism and long-distance dependencies, and the decoder carries out reconstruction operations on the potential representations to decode them into new music, which in turn makes the generated music closer to real music.

2024 4th International Conference on Electronic Information Engineering and Computer Science (EIECS) | 979-8-3315-3140-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/EIECS63941.2024.10800682

II. METHODS

A. Data pre-processing

In this paper, experiments are performed on the MAESTRO dataset [10]. Version 3.0 of this dataset contains a total of 1,276 MIDI sequences of piano performance works, MIDI is a music digitisation format in which note characterisation information can be easily extracted with music21.

In this paper, music is treated as a sequence of notes [11]. Note Sequence Representation is a method of representing music data as a sequence of notes. In MIDI format, each note is represented as an event, which contains information such as the pitch of the note, the start time, the end time and the duration of the note, etc. The specific structure is shown in Figure 1. The information about MIDI notes is encoded using the 'note on' marker, 'note off' marker and 'pitch' marker. In addition, the 'time shift' marker is used to indicate the relative time interval (in milliseconds) between two markers.

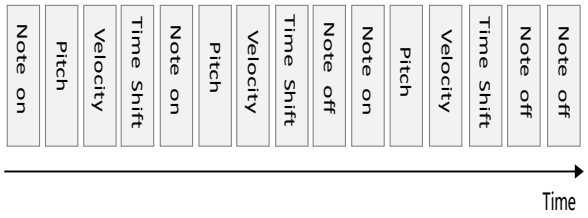

Figure 1. MIDI representation

By organising the information of each note into a vector, we can get a vector of note sequences for the whole piece of music, which is expressed as follows.

$$\text{tone sequence} = \left\{\left(p_1,\mathrm{n}_1,\mathrm{t}_1,\mathrm{v}_1\right),\ldots,\left(\mathrm{p}_N,\mathrm{n}_N,\mathrm{t}_N,\mathrm{v}_N\right)\right\} \quad (1)$$

Where, p_i denotes the pitch of the ith note, n_i denotes the start time of the ith note, t_i denotes the duration of the ith note, and v_i denotes the volume of the ith note. N is the length of the note sequence, indicating the number of notes contained in the music. With this representation, the music data can be converted into a sequence for processing and learning by the model.

B. Transformer

Transformer is a neural network model based on the attention mechanism, which has achieved great success in the field of natural language processing and is gradually being applied to other fields such as computer vision, music generation, etc. Transformer is an encoder and decoder structure, where an encoder consists of several encoding modules stacked together, and similarly, the decoder consists of the stacking of several layers of decoding modules Similarly, the decoder is a stack of multiple layers of decoding modules, as shown in Figure 2, and the number of layers of encoding and decoding modules can be set by ourselves during the design of the actual network architecture.

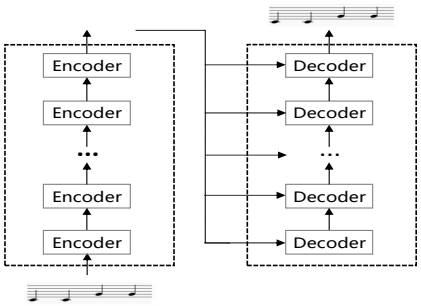

Figure 2. Transformer from encoder to decoder

The core of the Transformer model is the Multi-Headed Attention mechanism (MHA), which consists of a combination of multiple single-head attentions that are able to learn long-term dependencies in a sequence and extend the attention to different subspaces to capture data features from multiple perspectives. In the encoder and decoder, the output of the multi-head attention is further learnt through residual connectivity layers, layer normalisation, and feed-forward neural network layers.

C. Improved Transformer

The improved Transformer architecture diagram is shown in Figure 3.

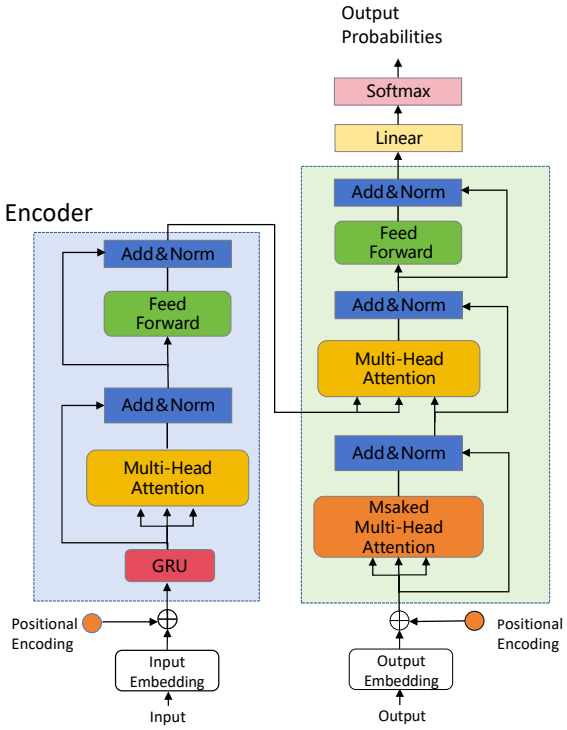

Figure 3. Diagram of the improved Transformer architecture

In order to capture the structure of music at a deeper level, this paper improves on the encoder part of Transformer by incorporating the GRU network into the encoder part of Transformer. Transformer performs well in sequence modelling, but the traditional LSTM (Long Short-Term Memory Network) still has its unique advantages in capturing

long-term dependencies. GRU is an optimised version of LSTM that improves training efficiency by simplifying the internal structure while maintaining almost the same effect. The unit structure diagram of GRU is shown in Figure 4.

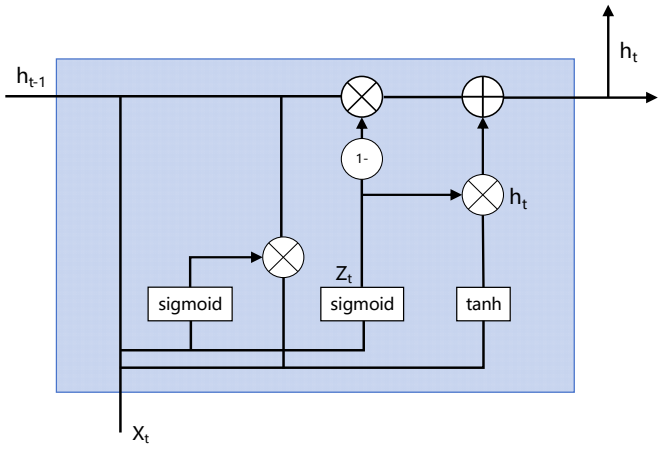

Fig. 4 Structure of GRU unit

The formula for each GRU unit is as follows:

(1) Reset door

$$r_t = \sigma\left(w_r[h_{t-1}, x_t]\right) \tag{2}$$

(2) Renewal Gate

$$z_t = \sigma\left(w_z[h_{t-1}, x_t]\right) \tag{3}$$

(3) Candidates for hidden states

$$\tilde{h}_t = \tanh\left(w[r_t \otimes h_{t-1}, x_t]\right) \tag{4}$$

(4) Final hidden state

$$h_t = (1 - z_t) \otimes h_{t-1} + z_t \otimes h_t \tag{5}$$

where x_t is the synchronization at time step t; h_{t-1} is the hidden state at the previous time t; σ is the Sigmod function; tanh activation function; and w_* is the weight matrix.

The input sequence first passes through the GRU network, and the sequence data are initially processed through the GRU to extract and enhance the time-dependent information in the input sequence to obtain the input vectors $X \in R^{d_h \times L}$ which can be obtained to three different matrix vectors through the linear transformation, which are the query matrix $Q \in R^{d_k \times L}$, the key matrix $K \in R^{d_k \times L}$, and the value matrix $V \in R^{d_k \times L}$, respectively. The linear transformation is shown in Equation 6.

$$Q = XW^Q, K = XW^K, XW^V \tag{6}$$

Where, $W^{Q,K,V} \in R^{d_h \times d_k}$ is the weight matrix, L is the sequence length, d_h is the hidden dimension of the model, and d_k is the hidden dimension of a single attention block.

After linear transformation, the attention layer aggregates the Q matrix and the corresponding K matrix, and then aggregates the obtained result with the V matrix to obtain the attention scores for each group of attention heads. The way to calculate the single group attention score is shown in Equation 7.

$$Z_i = Attentian(Q, K, V) = Soft\max(\frac{QK^T}{\sqrt{d_K}})V \tag{7}$$

Under the single-head attention mechanism, the model focuses on one or more specific locations without simultaneously affecting attention to other equally important locations. Therefore, MHA gives different subspaces to the attention layers. Specifically, different attention heads different attention heads use different Q, K and V matrices which, due to random initialisation, can project the trained input vectors into different subspaces and are processed by multiple independent attention heads in parallel, and finally the outputs of each attention head are aggregated to obtain a global attention score. The process of multi-head attention mechanism is shown in Equation 8.

$$Z = MutiAttentian(Q, K, V) = Concat(Z_1, Z_2, ..., Z_H)W^o \tag{8}$$

where H is the number of heads of attention and $W^o \in R^{d_h \times d_h}$ is the weight matrix.

III. EXPERIMENTS AND ANALYSIS OF RESULTS

In order to assess the effectiveness of the model proposed in this paper, experimental comparisons are mainly carried out in this study with the aim of comprehensively assessing and analysing the performance of the model from different perspectives.

A. Objective evaluation of music

The objective evaluation mainly includes the quantitative evaluation of the music generation model and the generated music. The strengths and weaknesses of the model are evaluated by comparing the data differences between the generated samples and the original samples. In order to more objectively evaluate the authenticity and quality of the music samples generated by the music generation network, this paper selects objective indicators with strong correlation with the music: Scale Consistency, Uniqueness, Tone Span, and Empty beat rate, and judges the validity of the music generated by the model by comparing it with other music generation models. The validity of the music generated by the model is judged by comparing it with other music generation models.

Scale Consistency, SC) is the percentage of notes in a sample of a counting music sequence that can best match the scale; Uniqueness, UN) is the percentage of pitches in the sample that are used only once; and Tone Span, TS) is the number of semitone steps between the lowest and the highest tones in the sample; Empty beat rate (EBR) is the ratio of beats played without any notes or instruments to the total number of beats in a rhythm.

In order to validate the effectiveness of the model in this paper, a variety of representative and commonly used music generation models are used for performance comparison. Table 1 shows the comparison results of this paper's model with Music VAE model (MusicVAE is a music generation model based on variational self-encoder proposed by Google's Magenta Studio), baseline Transformer model and Transformer-XL model (Transformer-XL adds a recursive mechanism on top of Transformer architecture) and real music in terms of various metrics. a recursive mechanism to the Transformer architecture) and comparison results of real music on various metrics.

Table 1. Comparison of objective evaluation indicators

Model	SC	UN	TS	EBR
Music VAE	79.3%	52	31	10.8%
Transformer	83.9%	50.4	23.6	7.76%
Transformer-XL	87%	57	18	6.4%
Our	90.1%	61.7	13.7	4.22%
Real Music	92.7%	64	14	2.56%

The average value of scale consistency of the music generated by this paper is close to the average value of scale consistency of the real music, which is better than the other comparative models, the uniqueness and pitch span perform well, the empty beat rate is lower than the other models, and the music is more coherent. On the whole, the quality of the music generated in this paper is better than the comparison models, and the generated music is closer to the real music in the sample set.

B. Subjective evaluation of music

The objective index of music can only be described from a quantitative point of view, music, after all, is to serve the human hearing, so the quality of a piece of music ultimately depends on the listener's evaluation of it, so far, the computer is currently not capable of evaluating the deep emotions and art embedded in the music, so it can't be completely simple to judge the music by objective evaluation. Therefore, in order to assess the perceived quality of the generated results more intuitively, this paper also designed a listening experiment to evaluate the music generated by different models. In this experiment, we invited 20 volunteers (10 men and 10 women, of which 8 were music lovers and the rest were general listeners) to participate in this experiment, and the volunteers were required to evaluate real music and music generated by previous models in addition to scoring the music pieces generated by this model.

In this paper, five pieces of music were randomly selected from the generation sample set of each model and the real music sample set. To ensure the objectivity and validity of the experiment, each piece of music is randomly disrupted and anonymously given to volunteers for evaluation and scoring. The music was scored on four dimensions: Harmoniousness (H), Consistency (C), Naturalness (N), and Overall-quality (OQ), using a five-point scale from 1 to 5, with 1 indicating very non-compliant, 2 indicating non-compliant, 3 indicating moderately compliant, 4 indicating more compliant, and 5 indicating very compliant, with higher scores representing better results. The final scores were averaged over the scores given by 20 people. The final results of the assessment are shown in Table 2.

Table 2. Results of the manual assessment

Model	H	C	N	OQ
Music VAE	2.6	3.4	3	2.5
Transformer	3	3.2	2.7	3
Transformer-XL	3.5	3.5	3.2	3.7
Our	4.1	3.7	3.8	4
Real Music	4.4	4.2	4	4.5

The subjective evaluation of the music samples generated by this paper's model is higher than that of other models, and the above subjective and objective evaluations show that this paper's model achieves good results in music composition, and can generate more harmonious and melodious music that is closer to real music.

IV. CONCLUSIONS

In this paper, a music generation model based on an improved Transformer is proposed for the field of intelligent music generation. Specifically, the gated recurrent unit (GRU) is used in the encoder of the Transformer, and the input sequence is processed using the GRU to capture longer dependencies, so that the attention can be focused on the global information, while the GRU provides an effective way to further capture local dependencies, which can further optimise the model performance. The experimental results show that the model proposed in this paper generates music that is closer to real music and strong audible compared to existing representative algorithms, which provides a new feasible idea for the direction of music generation.

REFERENCES

[1] Alpern A. Techniques for algorithmic composition of music[J]. On the web: http://hamp. hampshire.edu/adaF92/algocomp/algocomp, 1995, 95(1995): 120-120.

[2] Elliot Waite et al. Generating long-term structure in songs and stories. Magenta Bolg: https: //magenta.tensorflow.org/blog/2016/07/15/lookback-rnn-attention-rnn/, 2016.

[3] P. Tiwari and S. Jha, "Music Generation with Long Short-Term Memory Networks from MIDI Data of Classical Music,"2024 IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems (ICITEICS), Bangalore, India, 2024, pp. 1-4, doi: 10.1109/ICITEICS61368.2024.10625468.

[4] W. Liu, ‘‘Literature survey of multi-track music generation model based on generative confrontation network in intelligent composition,’’ J. Supercomput., vol. 79, no. 6, pp. 6560–6582, Apr. 2023.

[5] S. Sharma and S. Bvuma, ‘‘Generative adversarial networks (GANs) for creative applications: Exploring art and music generation,’’ Int. J. Multidisciplinary Innov. Res. Methodol., vol. 2, no. 4, pp. 29–33, 2023.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, ‘‘Attention is all you need,’’ in Proc. Adv. Neural Inf. Process. Syst., vol. 30, 2017, pp. 1-11.

[7] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Ian Simon, Curtis Hawthorne, Noam Shazeer, Andrew M Dai, Matthew D Hoffman, Monica Dinculescu, and Douglas Eck. Music transformer: Generating music with long-term structure. In International Conference on Learning Representations,2018.

[8] Y.-S. Huang and Y.-H. Yang, ‘‘Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions,’’ in Proc. 28thACM Int. Conf. Multimedia, Oct. 2020, pp. 1180–1188.

[9] N. Zhang, ‘‘Learning adversarial transformer for symbolic music generation,’’ IEEE Trans. Neural Netw. Learn. Syst., vol. 34, no. 4, pp. 1754–

1763, Apr. 2023

[10] Hawthorne, C., Stasyuk, A., Roberts, A., Simon, I., Huang, C.A., Dieleman, S., Elsen, E., Engel, J., & Eck, D. (2018). Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset. ArXiv, abs/1810.12247.

[11] Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. North American Chapter of the Association for Computational Linguistics.