IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 28 May 2023, accepted 13 June 2023, date of publication 19 June 2023, date of current version 28 June 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3287271

RESEARCH ARTICLE

Motif Transformer: Generating Music With Motifs

HENG WANG[1], **SEN HAO**[1], **CONG ZHANG**[2], **XIAOHU WANG**[1], **AND YILIN CHEN**[3]
[1]School of Mathematics and Computer, Wuhan Polytechnic University, Wuhan 430048, China
[2]School of Electrical and Electronic Engineering, Wuhan Polytechnic University, Wuhan 430048, China
[3]Hubei Key Laboratory of Intelligent Robot, Wuhan Institute of Technology, Wuhan 430073, China

Corresponding author: Sen Hao (13253620681@163.com)

This work was supported in part by the National Natural Science Foundation under Grant 61401319 and Grant 61272278; in part by the Hubei Province Natural Science Foundation under Grant 2014CFB270 and Grant 2015CFA061; in part by the Hubei Provincial Department of Education Research Foundation under Grant D20201601; in part by the Hubei Provincial Major Science and Technology Special Projects under Grant 2018ABA099; and in part by the Hubei Key Laboratory of Intelligent Robot, Wuhan Institute of Technology, under Grant HBIR202101.

ABSTRACT Music is composed of a set of regular sound waves, which are usually ordered and have a large number of repetitive structures. Important notes, chords, and music fragments often appear repeatedly. Such repeated fragments (referred to as motifs) are usually the soul of a song. However, most music generated by existing music generation methods can not have distinct motifs like real music. This study proposes a novel multi- encoders model called Motif Transformer to generate music containing more motifs. The model is constructed using an encoder-decoder framework that includes an original encoder, a bidirectional long short term memory-attention encoder (abbreviated as bilstm-attention encoder), and a gated decoder. Where the original encoder is taken from the transformer's encoder and the bilstm-attention encoder is constructed from the bidirectional long short-term memory network (BILSTM) and the attention mechanism; Both the original encoder and the bilstm-attention encoder encode the motifs and input the encoded information representations to the gated decoder; The gated decoder decodes the entire input of the music and the information passed by the encoders and enhances the model's ability to capture motifs of the music in a gated manner to generate music with significantly repeated fragments. In addition, in order to better measure the model's ability of generating motifs, this study proposes an evaluation metric called used motifs. Experiments on multiple music field metrics show that the model proposed in this study can generate smoother and more beautiful music, and the generated music contains more motifs.

INDEX TERMS Deep learning, music generation, recurrent neural network, transformer.

I. INTRODUCTION

Music is an organized and regular sound wave, which can cultivate the mood and relax, convey messages, and express emotions. It is a necessary art form in our daily life. However, composing or analyzing music requires a high degree of professionalism, and it takes inspiration and various human and material resources for music experts to complete a song. In order to better analyze and study music, many scholars use artificial intelligence technology for music-related tasks such as emotion recognition [1], music classification [2], and music generation [3]. Composing music using artificial intelligence can be automated or semi-automated using neural network models to generate music, significantly reducing the complexity of managing music.

In recent years, sequence-based music generation methods have made significant progress with the development of deep learning techniques. Sequence-based methods [4], [5] will first quantize the music content into symbolic sequences and then input the symbolic sequences into a neural network model. As shown in Fig.1, score fragments of the Christmas pop song "Deck the Halls" consist of notes as the basic unit and are divided into different parts by bars [6]. Additional notes have their pitch, playing time, and duration.

The associate editor coordinating the review of this manuscript and approving it for publication was Angel F. García-Fernández.

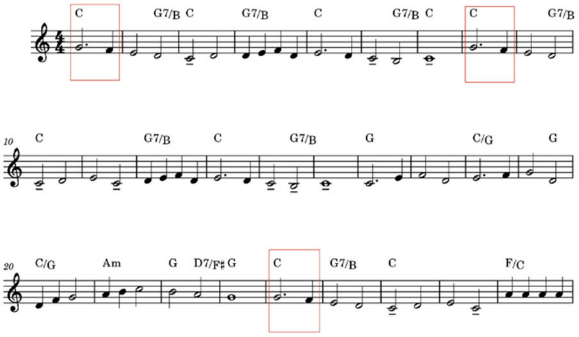


FIGURE 1. The score fragments of "Deck the Halls".

The sequence-based approach will quantify the above musical information, such as notes, bars, and rhythms, into symbolic sequences for model learning. In recent years, many scholars have made good progress in music generation using neural network models such as recurrent neural networks [7], [8], generative adversarial networks [9], [10], and variational autoencoders [11]. Music usually has obvious regularity in its overall structure, and a standard piece of music typically contains many repetitive fragments that run through the entire structure of the music rather than just being reflected in the short term. For example, the musical fragments in the red box in Fig.1 recur several times throughout the song. Such recurring melodic fragments (motifs) are often impressive and key to a song's tone. However, we found that the above music generation methods can only contain some repetitive fragments in the first few bars of the generated music rather than the entire music having repetitive segments.

Therefore, this study proposes a music generation model called Motif Transformer that can generate more motifs. Specifically, this study designed a transformer-based model containing an original encoder, a bilstm-attention encoder, and a gated decoder. The model enhances the understanding of the encoding of motifs through original and bilstm-attention encoders and improves the model modeling of crucial information from the encoders through gated decoders, thus strengthening the model's focus on motifs to generate music with impressive features. This study designed experiments to compare the generation performance of the Motif Transformer with other music generation models [12], [13]. And we developed objective and human subjective evaluation metrics to assess the gap between generated and authentic music.

The rest of this article is organized as follows. Section II briefly discusses the relevant work. Section III introduces the details of the proposed method. Section IV provides implementation details and experimental results. Finally, Section V provides a summary of the paper and discusses the potential contributions and future development directions of the method proposed in this study.

II. RELATED WORK

The composition of music by algorithms on computers dates back to the 1950s. After the first computers were invented and built, mathematician and composer Hiller used Markov chains in combination with knowledge of music theory to compose the first computer-generated music, pioneering the creation of music by artificial intelligence. Subsequently, with the development of deep learning, many scholars began to study using neural network models for music composition. For example, Casella et al. [7] proposed the melody_RNN method for music generation through a single-layer long and short-term memory network (LSTM) and a fully connected layer, which can combine current and historical music information to generate new music melodies. However, the generated music does not have long-time structural connections due to the gradient disappearance problem of recurrent neural networks. Hadjeres et al. [14] proposed the Deep bash method. Unlike traditional recurrent neural networks, Deep bash generates music along the time axis but selects a time to generate from the middle to both sides, alleviating the long-term dependency problem of traditional recurrent neural networks. Keerti et al. [15] combined the bidirectional long short-term memory network with the attentional mechanism to generate jazz music with a repetitive structure. The above studies alleviate the gradient disappearance problem of recurrent neural networks by various methods, hoping to generate music containing more motifs using recurrent neural networks. Still, as of now, recurrent neural network-based models are only excellent at generating short segments of music, and the modeled information is difficult to establish long structural connections.

Many scholars have also employed other networks to solve the problem of difficulty in modeling long-structured linked music. For example, Guan et al. [9] and Arora et al. [10] used generative adversarial networks to generate music and Grekow et al. [11] for modeling musical information using variational autoencoders. However, their achievements are still unsatisfactory. It was not until Vaswani et al. [16] proposed the Transformer model which made a sensation in artificial intelligence. Transformer is a self-attention-based sequence model with strong self-attention and modeling capabilities that have achieved excellent results in many natural language processing tasks [17], [18]. Shortly after the release of Transformer, Huang et al. [12] proposed the Music Transformer to pioneer the use of Transformer for modeling music generation tasks, and they used Transformer's powerful ability to handle long sequences to generate more structurally connected music melodies; Subsequently, Hsiao et al. [19] proposed the Compound Word Transformer, which represents musical events in the form of compound words, significantly reducing the length of musical sequences and speeding up model convergence while generating music of comparable quality; Choi et al. [20] proposed a chord-conditioned Melody Transformer, which uses a transformer to generate rhythms and pitches conditional on chords. The model can effectively learn the pitch and rhythm distribution of music, and there are some significant repetitive fragments in the generated music melody. Shih et al. [13] proposed the Theme Transformer, which extracts repetitive fragments of musical melodies as thematic material for the input model and makes the model focus more on thematic fragments by a novel position encoding method and a method for balancing attention mechanisms. It is possible to generate significant repetitive fragments for short and medium-sized music. However, the performance of generating repetitive fragments on long music still needs to be improved.

Due to the powerful modeling capability of Transformer and its excellent performance in maintaining long-term structural consistency, this study proposes the Motif Transformer model based on Transformer. And inspired by [21] and [22], our proposed model contains multiple encoders. Specifically, the Motif Transformer includes an original encoder, a bilstm-attention encoder, and a gated decoder; original and bilstm-attention encoders encode motifs and pass the encoded information to the gated decoder through their respective cross-attention mechanisms; The gated decoder balances cross-attention and self-attention through gate control, allowing the model to capture motif information better and generate music containing more motifs.

III. APPROACH

In this section, this study describes the proposed model named Motif Transformer. Sequence-based music generation methods typically convert the music in midi format into symbol sequences and input them into the model. Each note in music requires multiple symbol representations, which leads to a significant increase in sequence length and complexity. Therefore, traditional Transformer based music generation methods may have the following two drawbacks when generating music. Firstly, due to the lack of ability to capture temporal information in the Transformer model's self-attention mechanism, the encoder of the model is difficult to capture temporal information in the sequences. Secondly, due to the long and complex sequence, the decoder may ignore the motif information transmitted by the encoder. To this end, we propose two methods to enhance the model's ability to generate motifs. Firstly, we used a multi encoders architecture to improve the model's encoding ability [21], [22]. We have designed an encoder called bidirectional long short term memory-attention encoder (bilstm-attention encoder) based on the bidirectional long short term memory network (BILSTM) and the attention mechanism. The bilstm-attention encoder can capture temporal information in music sequences, making up for the shortcomings of the self-attention mechanism and providing more information about motifs for the model. Secondly, we designed a gated decoder to generate more motifs. In our model, the encoders are responsible for encoding motifs and passing the encoded information to the decoder through the cross-attention mechanism. The decoder receives the entire music sequences and information from the encoders and generates music based on them. To enhance the model's ability to generate motifs, we have designed a gating mechanism to turn off some layers of self-attention mechanism, so that the decoder pays more attention to the music motif information encoded by the encoders through the cross-attention mechanisms. In the motif areas, we use cross-attention mechanism in all layers, while only using self-attention mechanism in the first two layers. In non-motif areas, all layers use self-attention. This makes it easier for the model to interact with the motif information transmitted by the encoders through the cross-attention mechanisms, thereby generating music containing more motifs.

As shown in Fig.2, our model contains three main modules, the original encoder, the bilstm-attention encoder, and the gated decoder. The original and the bilstm-attention encoders model the motifs of music and then add them to the gated decoder via the cross-attention mechanism. The gated decoder receives the complete music sequences and the information from the encoders and generates the music content. In the following, this study will describe the encoders and decoder in the model separately.

A. ORIGINAL ENCODER

The role of the Transformer encoder is to encode the input sequences into an internal representation. It contains multiple attention layers, with each layer encoding the previous layer's output. Each attention layer has a self-attention mechanism, which calculates weights for the hidden states at each location. These weights show how well the hidden state at the current location is related to the hidden conditions at other locations. The attention mechanism allows the model to extract information from the entire sequences and encode them into the internal representation. In general, a Transformer model uses only one encoder, but using multiple independent encoders can improve the expressiveness of the model and allow the model to capture better the complex relationships in the sequences [21], [22]. Based on this, this study designed two encoders for our model: the original encoder and the bilstm-attention encoder, and we will introduce them separately below.

Transformer has strong modeling and long-term structural consistency capabilities, and its encoder has strong encoding capabilities. Many music generation methods use Transformer's encoder and have achieved excellent performance [12], [13]. This article retains the Transformer's original encoder as one of the model's encoders and names it the original encoder. As shown in the left side of Fig.2, the original encoder consists of an embedding layer, a position encoding, and an attention layer, where the attention layer is a stack of multi-headed attention and feedforward networks. Suppose the input music clips are $X = \{x_1, x_2 \cdots x_\tau\}$, then the embedding encoding and position encoding can be expressed as:

$$XV = Emb\,(X) + PositionEng\,(X) \tag{1}$$

After being encoded, X is fed into the attention layer. The attention layer is the core of the traditional encoder and contains a multi-headed attention mechanism and a feed-forward neural network. After entering the attention layer, X is first transformed linearly into Q, K, and V, to get the information

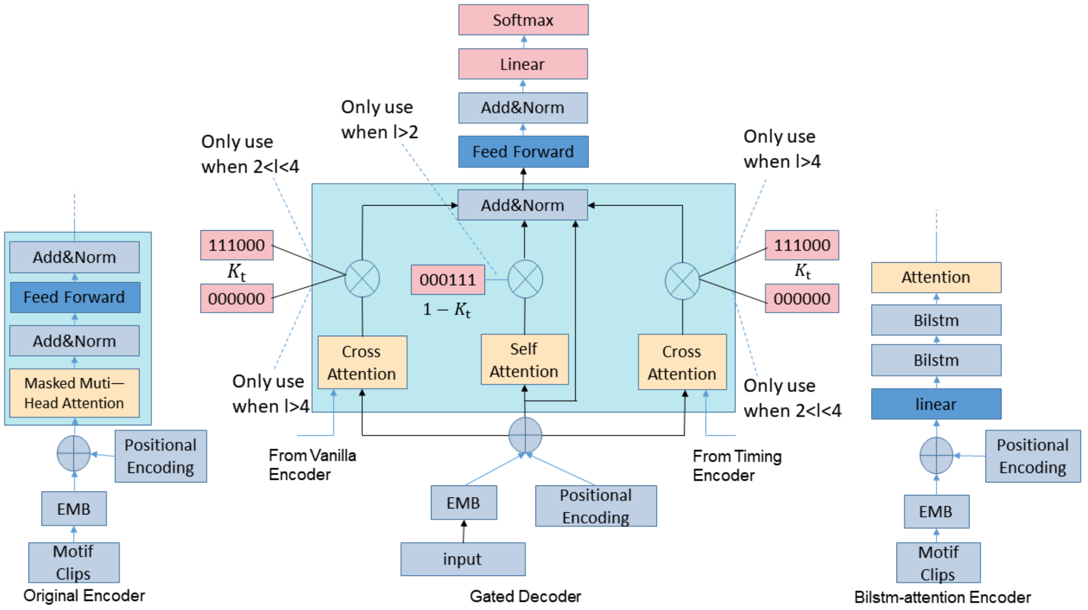


FIGURE 2. Model architecture diagram. On the left side of the figure is the original encoder; in the middle of the figure is the gated decoder; and on the right side of the figure is the bilstm-attention encoder.

of different spatial locations of X itself, as shown in (2).

$$Q, K, V = X_V W_Q, X_V W_K, X_V W_V \tag{2}$$

Next, Q, K, and V are again linearly transformed to input multiple attention mechanisms. After that, multiple attentions are connected to get the output of multiple attentions, as shown in (3) (4).

$$headh = Attention\left(QW_h^Q, KW_h^K, VW_h^V\right) \tag{3}$$

$$MultiHeadAttn\,(Q, K, V) = Concat\,(head_1, \cdots, head_H)\, W^o \tag{4}$$

where *Attention*() represents the attention mechanism, *MultiHeadAttn*() represents the multi-head attention function, *Concat*($head_1$,...,$head_H$) represents the output of multiple attentions stitched together, and $head_H$ represents the output of a single attention.

The output of multi-headed attention is followed by a fully connected feedforward neural network, and there are residual connections and normalization calculations at both the input and output of the feedforward neural network, which are shown in the following (5) (6) (7).

$$Lm = LayerNorm\,(A + X) \tag{5}$$

$$P = Position - wise - Feed - Forward\,(Lm) \tag{6}$$

$$O = LayerNorm\,(P + Lm) \tag{7}$$

where *LayerNorm*() denotes the regularization process; *Position-wise-Feed-Forward*() denotes the feed-forward network calculation.

B. BILSTM-ATTENTION ENCODER

Transformer's self-attention mechanism can model the association between the current position and other positions at once, and has good global modeling ability. However, the self-attention mechanism cannot capture the temporal information of the sequences. Because the bidirectional long short term memory network has a better ability to capture time information, this study additionally designs an encoder named bidirectional long short term memory-attention encoder (bilstm-attention encoder), which consists of an embedding layer, a position encoding layer, a linear layer, two bidirectional long short term memory networks, and a attention mechanism. In this way, the bilstm-attention encoder will combine the global modeling ability of the attention mechanism with the time information modeling ability of BILSTM. Like the original encoder, the bilstm-attention encoder also encodes only motifs. Still, unlike the original encoder, which can only encode self-internal connections between messages, the bilstm-attention encoder can model the relationships of input information at the time scale, bringing temporal information to the model and enhancing the model's understanding of motifs. Assuming that the input music fragments are $X = \{x_1, x_2 \cdots x_\tau\}$, the computation of the embedding layer and position encoding of the bilstm-attention encoder can be expressed as in (8).

$$X_T = Emb\,(X) + PositionEng\,(X) \tag{8}$$

Next, the music fragments after the computed embedding encoding and location encoding are fed into the linear layer, calculated as in (9).

$$X_L = Linear\,(X_T) \tag{9}$$

After the linear transformation, we want the next layer of the computation process to contain both past and future music fragments information. Therefore, we use a two-layer BiLSTM. The BiLSTM can use a forward LSTM to obtain forward implied state information and a backward LSTM to obtain backward implied state information [23].

$$L_1 = Bilstm\,(X_L) \tag{10}$$

$$L_2 = Bilstm\,(L_1) \tag{11}$$

Finally, to enhance the ability of the model to establish structural connections, we set up a layer of attention mechanism layer, which is calculated as in (12).

$$A_T = Attention\,(L_2) \tag{12}$$

C. GATED DECODER

The role of the decoder is to decode the output of the encoder and generate the target sequences. The decoder contains multiple concatenated attention layers that decode the input information. Unlike the encoder, the decoder uses two attention mechanisms in its modeling process: self-attention and cross-attention. Self-attention provides an internal understanding of the relationship between the current hidden state and the sequence generated during decoding, with strong self-attention and local dependence, whereas the cross-attention mechanism interacts with the information in the encoder with that in the decoder, enhancing the model's understanding of information about different modalities or different spatial locations. Balancing these two attention mechanisms allows the model to take full advantage of their respective strengths and improves the model's performance [24].

This study places self-attention and two cross-attention mechanisms in parallel and designs a gating mechanism to control the information flow between self-attention and cross-attention mechanisms. As shown in the middle of Fig.2, in the first two layers of the decoder, this study uses self-attention mechanism in all areas, while only cross-attention mechanisms are used in the motif areas; In the middle two layers of the decoder, this study uses self-attention only for non-motif areas and uses cross-attention linked to the original encoder in the motif areas; In the last two layers of the decoder, this study uses self-attention only for non-motif areas and uses cross-attention connected to the bilstm-attention encoder in the motif areas. If we let f_t^l be the output of the lth decoder at time step t, the mathematical equation for the process is expressed as follows.

$$f_t^l = \begin{cases} k_t f_t^{l,(cross1)} + (1-k_t)\, f_t^{l,(self)}, & l > 4 \\ k_t f_t^{l,(cross2)} + (1-k_t)\, f_t^{(l,(self))}, & 2 < l < 4 \\ k_t f_t^{l,(cross1)} + k_t f_t^{l,(cross2)} + f_t^{l,(self)}, & l < 2 \end{cases} \tag{13}$$

where k_t indicates whether the sequence is within the motifs at time step t. If it is within the motifs then $k_t = 1$, otherwise $k_t = 0$. $f_t^{l,(cross1)}$ denotes the output of cross-attention connected to the conventional encoder, $f_t^{l,(cross2)}$ denotes the output of cross-attention connected to the temporal encoder, and $f_t^{l,(self)}$ denotes the output of self-attention.

IV. EXPERIMENTS

A. DATASET AND DATA PREPROCESSING

This study uses the POP909 dataset [25] to perform experiments on our model. The POP909 dataset contains 909 popular music tracks in midi form composed by professional musicians. The songs in the dataset include three main types of musical information: the main melody (with the vocal pair thereof), the secondary melody, and the piano accompaniment. In this case, the secondary melody plays a minor role in the whole song, and its removal does not affect the quality of the song too much [13]. Referrin to [19], this study used information from only two tracks, the main melody and the piano accompaniment, and selected the songs in 4/4 time in them. After the selection was completed, a total of 713 songs were available for the experiment, of which we took 29 songs as the test set, and the remaining songs were used for model training.

To make the POP909 dataset available for experiments on the model, we need to extract music features from music files in midi format and combine them into event sequences. This study uses two tracks (main melody and piano accompaniment) from the POP dataset songs, and the extracted note- related sequences are Note-On-Melody, Note-Duration-Melody, Note-Velocity-Melody, Note-On-Piano, Note-Duration-Piano, and Note-Velocity-Piano; Rhythm and position related sequences as Tempo, Bar, and Position; in which, Note-On indicates the pitch start time (range is 1-127), Note-Duration indicates the pitch playing duration (1/4 beat as the basic the unit, range is 1-64), Note-Velocity indicates the pitch playing intensity (range is 1-126), Tempo means the song playing tempo, taking the value of 17bpm-194bpm, Bar shows the number of bars, and Position indicates the position of each event. In addition, following [13], motif-start and motif-end are set in the music motifs to mark the start and end of the music motifs. Fig.3, for example, shows the conversion of midi music into an event sequence.

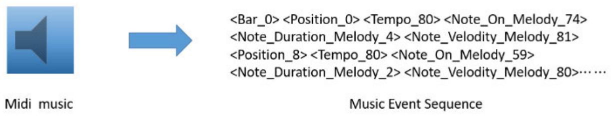

Midi music
<Bar_0> <Position_0> <Tempo_80> <Note_On_Melody_74>
<Note_Duration_Melody_4> <Note_Velodity_Melody_81>
<Position_8> <Tempo_80> <Note_On_Melody_59>
<Note_Duration_Melody_2> <Note_Velodity_Melody_80>······
Music Event Sequence

FIGURE 3. Example of converting Midi music into a sequence of music events.

B. MODEL SETTING

In the experiment of this study, we set the maximum length of the sequence of music events to 512. The bilstm-attention encoder of the model is trained through the Bi LSTM network with a hidden layer of 256. Both the original encoder and the gated decoder of the model consist of 6 attention layers, each with eight attention heads, and a feedforward neural network hidden dimension of 1024. The configuration of our experimental environment is shown in Table.1. Referring to [13], the experiment respectively selected 29 songs (about 4%) from a dataset of 713 songs as the test set and validation set, and used the remaining songs as the training set. This

TABLE 1. Experimental environment table.

Category	Version
CPU	Intel(R) Core (TM) i7-6700K CPU@4.00GH
GPU	NVIDIA GeForce RTX 3090
Memory	24GB
Python	3.6

experiment used Adaptive Moment Estimation for optimization and cross entropy as the loss function. Set the learning rate of the Adam optimizer to 0.0001 and the number of iterations for training to 2000.

C. OBJECTIVE EVALUATION

POP909 dataset was used for this evaluation to train Attentional network [15], Music Transformer [12] and Theme Transformer [13], and their results were compared with the Motif Transformer. In terms of objective metrics selection, this study first selected the null beat rate [26] and pitch class entropy [27] to objectively evaluate the generated music and compared the music generated by the models with the real music, and the closer the generated music is to the real music, the better the performance of the model. In addition, this study also designed an objective metric to measure the model's ability to generate motifs: used motifs (UM). The higher metric value indicates the model's stronger ability to generate motifs. Each of the three metrics is described below.

Empty beat rate (EBR) is the ratio of beats played without any notes or instruments to the total number of beats in a rhythm [26]. Empty beat rate is often used to measure the sense of empty inspiration or emptiness in a piece of music. Also known as the "rest ratio", the airtime ratio is a standard metric used in music analysis and generation. Define (14), where *empty_beat* denotes the number of empty beats and *all_beat* indicates the number of all beats.

$$Empty\_beat\_rate = \frac{empty\_beat}{all\_beat} \qquad (14)$$

Pitch class entropy (PCE) can be used to describe the uniformity and complexity of the musical pitch distribution [27]. In the field of music, each note corresponds to a pitch level, and pitch entropy is the statistical quantity used to describe the distribution of these levels as they occur in music. A lower pitch entropy indicates a more regular and biased musical structure. Define (15), where *P(pitch=i)* denotes the probability of occurrence of the *i*-th category of pitches.

$$Pitch\_class\_entropy = -\sum_{i=0}^{11} P\left(pitch = i\right) \log_2 \left(P\left(pitch = i\right)\right) \qquad (15)$$

To verify the model's ability of generating motifs, we propose an objective metric called used motifs (UM). The repetitive fragments (also called motifs) in a piece of music are often used to express the tone and feel of the music and are the heart of a song. In this study, we measure the model's ability to generate motifs by counting the ratio of the number of motifs to the number of bars in the generated music. Define (16), where *motifs* denote the number of occurrences of motifs in the music, and *n_bars* denotes the number of bars.

TABLE 2. The entropy of null beat rate and pitch class for model-generated music as well as real music.

Model	EBR	PCE
Attentional Network [15]	0.95	2.86
Music Transformer [12]	0.91	2.75
Theme Transformer [13]	0.89	2.83
Motif Transformer	0.86	2.72
Real music	0.81	2.66

$$Used\_Motifs = \frac{motifs}{n\_bars} \qquad (16)$$

This study had each model generate ten pieces of music, then each calculated the EBR and PCE and averaged them, and compared their results with those of real music. To evaluate the performance of the Motif Transformer, we selected three most typical open-source models, namely:

1) Attentional network: A model combining attention mechanism and bidirectional Long short-term memory network can generate music with some repetitive structures [15].
2) Music Transformer: A music generative model based on the relative attention mechanism proposed by Google Brain can generate long-term music with motifs [12].
3) Theme Transformer: A new transformer-based model that proposes a novel position encoding method and a method for balancing attention mechanisms, specifically for generating music with motifs [13].

As shown in Table.2, our model achieves the minimum results and is closest to the real music for both objective metrics. This indicates that the music generated by the model proposed in this study has a better sense of rhythm and structural regularity than other models and is closer to the real music.

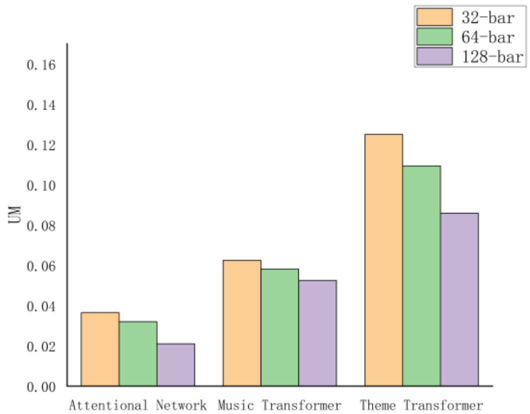


FIGURE 4. Distribution of the UM results, each model generates ten pieces of 32-bar, 64-bar, and 128-bar music, and the average UM of the ten pieces of music is shown in the figure.

Fig.4 shows the distribution of UM for each model. From it, we can find that the more bars of music generated by all models, the lower the relative UM values. This indicates that all models are less capable of generating long-structured music than short-structured music. In addition, compared with other models, our model generates music with higher UM values regardless of the number of bars, indicating that our model can generate motifs better.

D. SUBJECTIVE EVALUATION

The value and quality of music is a subjective feeling, and so far, it is not possible to judge the music entirely simply by objective evaluation. Therefore, this study also designed listening experiments to evaluate the music generated by the different models. In the experiment, 20 volunteers (10 of each gender, 5 of whom were professional music practitioners) rated musical works on a scale from 1 to 5 based on the following five indicators. Each model generates ten pieces of music for the 64 bars, and each volunteer selects two pieces from each model to listen to and evaluate. And to ensure the experiment's validity, the music after each volunteer's selection was randomly disrupted and then made available for volunteers to evaluate and score.

- Truth: Is this music consistent with human creative habits?
- Structure: Does this music have distinctly repetitive segments?
- Harmony: Does the melody of this music sound harmonious?
- Accuracy: Is this music free of compositional and performance errors?
- Pleasure: Does this music sound good and pleasant?

TABLE 3. Subjective evaluation results.

Model	T	S	H	A	P
Attentional Network [15]	3.3	3.2	3.3	3.5	3.5
Music Transformer [12]	3.5	3.3	3.5	3.7	3.6
Theme Transformer [13]	3.9	4.0	4.1	4.1	4.2
Motif Transformer	4.1	4.3	4.2	4.2	4.3

As can be seen from Table.3, the Motif Transformer proposed in this study has better performance in all human subjective evaluation metrics. This shows that Motif Transformer can generate harmonious and natural music. In particular, the score of structure has been improved compared with other models. This shows that Motif Transformer has a more significant advantage in generating motifs, validating the effectiveness of the model improvements.

V. CONCLUSION

This study proposes Motif Transformer for the feature that music has structural repetitiveness. This study designed a music generation method that combines multiple encoders and a gated decoder. Motif Transformer enhances the understanding of encoding of motifs through the original encoder and bilstm-attention encoder, and enhances the model modeling of motif information from encoders through the gated decoder, thus allowing the model to gain more attention to motif information. Moreover, this study proposes an objective metric called used motifs to verify the ability of the model to generate motifs and a subjective listening experiment to test the model's validity. After experimental verification, the network model proposed in this study can generate harmonious and natural music, and the generated music contains more motifs.

Compared with traditional music generation methods, the method proposed in this article can generate music with specific motifs. According to the method proposed in this article, people can obtain a large number of music works more conveniently, and the theme style of these music works can be selected according to personal needs, which can significantly promote music consumption. Moreover, the method proposed in this article can also provide more possibilities for those engaged in music creation and production, bringing them more creative and inspirational inspiration, thereby promoting the development of the music industry. In addition, the current model only performs well in generating one type of motif. A real song may contain different types of motifs. In the future, we will explore using the model to generate music that contains multiple types of motifs.

REFERENCES

[1] K. Markov and T. Matsui, "Music genre and emotion recognition using Gaussian processes," *IEEE Access*, vol. 2, pp. 688–697, 2014, doi: 10.1109/ACCESS.2014.2333095.

[2] M. Ashraf, G. Geng, X. Wang, F. Ahmad, and F. Abid, "A globally regularized joint neural architecture for music classification," *IEEE Access*, vol. 8, pp. 220980–220989, 2020, doi: 10.1109/ACCESS.2020.3043142.

[3] I. Goienetxea, I. Mendialdua, I. Rodríguez, and B. Sierra, "Statistics-based music generation approach considering both rhythm and melody coherence," *IEEE Access*, vol. 7, pp. 183365–183382, 2019, doi: 10.1109/ACCESS.2019.2959696.

[4] M. K. Jedrzejewska, A. Zjawinski, and B. Stasiak, "Generating musical expression of MIDI music with LSTM neural network," in *Proc. 11th Int. Conf. Hum. Syst. Interact. (HSI)*, Gdansk, Poland, Jul. 2018, pp. 132–138, doi: 10.1109/HSI.2018.8431033.

[5] S. Oore, I. Simon, S. Dieleman, D. Eck, and K. Simonyan, "This time with feeling: Learning expressive musical performance," *Neural Comput. Appl.*, vol. 32, no. 4, pp. 955–967, Feb. 2020, doi: 10.1007/s00521-018-3758-9.

[6] Y. Qin, H. Xie, S. Ding, B. Tan, Y. Li, B. Zhao, and M. Ye, "Bar transformer: A hierarchical model for learning long-term structure and generating impressive pop music," *Appl. Intell.*, vol. 53, no. 9, pp. 10130–10148, 2023, doi: 10.1007/s10489-022-04049-3.

[7] P. Casella and A. Paiva, "MAgentA: An architecture for real time automatic composition of background music," in *Proc. Int. Workshop Intell. Virtual Agents*, Berlin, Germany, pp. 224–232, 2001, doi: 10.1007/3-540-44812-8_18.

[8] D. Eck and J. Schmidhuber, "A first look at music composition using LSTM recurrent neural networks," Istituto Dalle Molle Di Studi Sull Intelligenza Artificiale, Tech. Rep. IDSIA-07-02, 2002, vol. 103, no. 4, pp. 48–56. [Online]. Available: https://dl.acm.org/doi/book/10.5555/870511

[9] F. Guan, C. Yu, and S. Yang, "A GAN model with self-attention mechanism to generate multi-instruments symbolic music," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Budapest, Hungary, Jul. 2019, pp. 1–6, doi: 10.1109/IJCNN.2019.8852291.

[10] S. Arora, A. Dassler, T. Earls, M. Ferrara, N. Kopparapu, and S. Mathew, "An analysis of implementing a GAN to generate MIDI music," in *Proc. IEEE MIT Undergraduate Res. Technol. Conf. (URTC)*, Cambridge, MA, USA, 2022, pp. 1–5, doi: 10.1109/URTC56832.2022.10002181.

[11] J. Grekow and T. Dimitrova-Grekow, "Monophonic music generation with a given emotion using conditional variational autoencoder," *IEEE Access*, vol. 9, pp. 129088–129101, 2021, doi: 10.1109/ACCESS.2021.3113829.

[12] C.-Z. Anna Huang, A. Vaswani, J. Uszkoreit, N. Shazeer, I. Simon, C. Hawthorne, A. M. Dai, M. D. Hoffman, M. Dinculescu, and D. Eck, "Music transformer," 2018, *arXiv:1809.04281*.

[13] Y. Shih, S. Wu, F. Zalkow, M. Müller, and Y. Yang, "Theme transformer: Symbolic music generation with theme-conditioned transformer," *IEEE Trans. Multimedia*, early access, Mar. 23, 2022, doi: 10.1109/TMM.2022.3161851.

[14] G. Hadjeres, F. Pachet, and F. Nielsen, "DeepBach: A steerable model for bach chorales generation," 2016, *arXiv:1612.01010*.

[15] G. Keerti, A. N. Vaishnavi, P. Mukherjee, A. S. Vidya, G. S. Sreenithya, and D. Nayab, "Attentional networks for music generation," *Multimedia Tools Appl.*, vol. 81, no. 4, pp. 5179–5189, Feb. 2022, doi: 10.1007/s11042-021-11881-1.

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 2, 2017, pp. 5999–6009.

[17] Q. Guo, J. Huang, N. Xiong, and P. Wang, "MS-pointer network: Abstractive text summary based on multi-head self-attention," *IEEE Access*, vol. 7, pp. 138603–138613, 2019, doi: 10.1109/ACCESS.2019.2941964.

[18] C. Wen and L. Zhu, "A sequence-to-sequence framework based on transformer with masked language model for optical music recognition," *IEEE Access*, vol. 10, pp. 118243–118252, 2022, doi: 10.1109/ACCESS.2022.3220878.

[19] W. Y. Hsiao, J. Y. Liu, and Y. C. Yeh, "Compound word transformer: Learning to compose full-song music over dynamic directed hypergraphs," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 178–186.

[20] K. Choi, J. Park, W. Heo, S. Jeon, and J. Park, "Chord conditioned melody generation with transformer based decoders," *IEEE Access*, vol. 9, pp. 42071–42080, 2021, doi: 10.1109/ACCESS.2021.3065831.

[21] A. Abdelraouf, M. Abdel-Aty, and J. Yuan, "Utilizing attention-based multi-encoder–decoder neural networks for freeway traffic speed prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 11960–11969, Aug. 2022, doi: 10.1109/TITS.2021.3108939.

[22] J. Bi, L. Zhang, H. Yuan, and J. Zhang, "Multi-indicator water quality prediction with attention-assisted bidirectional LSTM and encoder–decoder," *Inf. Sci.*, vol. 625, pp. 65–80, May 2023, doi: 10.1016/j.ins.2022.12.091.

[23] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Vancouver, BC, Canada, Mar. 2013, pp. 6645–6649, doi: 10.1109/ICASSP.2013.6638947.

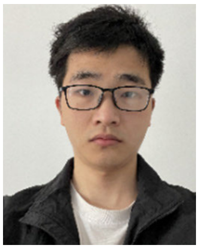
[24] H. Rashkin, A. Celikyilmaz, Y. Choi, and J. Gao, "PlotMachines: Outline-conditioned generation with dynamic plot state tracking," 2020, *arXiv:2004.14967*.

[25] Z. Wang, K. Chen, J. Jiang, Y. Zhang, M. Xu, S. Dai, X. Gu, and G. Xia, "POP909: A pop-song dataset for music arrangement generation," 2020, *arXiv:2008.07142*.

[26] W. H. Dong, W. Y. Hsiao, and Y. H. Yang, "Pypianoroll: Open source Python package for handling multitrack pianoroll," in *Proc. ISMIR*, Paris, France, 2018, pp. 120–127, doi: 10.5281/zenodo.4540221.

[27] S.-L. Wu and Y.-H. Yang, "The jazz transformer on the front line: Exploring the shortcomings of AI-composed music through quantitative measures," 2020, *arXiv:2008.01307*.

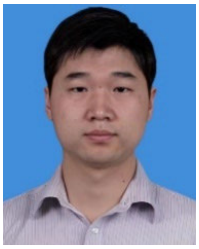
HENG WANG received the B.E. degree from the Huazhong University of Science and Technology, in 2006, and the Ph.D. degree in engineering from Wuhan University, in 2013. He is currently a Professor with the School of Mathematics and Computer Science, Wuhan Polytechnic University. He is also a Postdoctoral Research Fellow with Alto University, Finland. His research interests include the perception characteristics of acoustic spatial parameters, artificial intelligence, and the application of 3D audio and video in virtual reality.



SEN HAO received the B.E. degree from the Henan University of Technology, Zhengzhou, China, in 2021. He is currently pursuing the M.S. degree in software engineering with Wuhan Polytechnic University, Wuhan. His research interests include music information retrieval and artificial intelligence technology and its application.

CONG ZHANG received the bachelor's degree in automation engineering from the Huazhong University of Science and Technology, in 1993, the master's degree in computer application technology from the Wuhan University of Technology, in 1999, and the Ph.D. degree in computer application technology from Wuhan University, in 2010. He is currently a Professor with the School of Electrical and Electronic Engineering, Wuhan Polytechnic University. His research interests include multimedia signal processing, multimedia communication system theory and application, and pattern recognition.

XIAOHU WANG received the B.E. degree from Tianjin Chengjian University, Tianjin, China, in 2021. He is currently pursuing the M.S. degree in software engineering with Wuhan Polytechnic University, Wuhan. His research interests include music information retrieval and artificial intelligence technology and its application.



YILIN CHEN received the Ph.D. degree from the Department of Computer Science, Wuhan University, China, in 2020. He is currently a Lecturer with the School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan, China. His research interests include multi-objective optimization, intelligent optimization, image processing, and computer graphics.