Computer Fraud and Security
ISSN (online): 1873-7056

Research on Automatic Composition and safety of Chinese Guzheng Based on Joint Cross Attention

Sisi Zhu[1], Shimin Wang[2], Ming Zhao[34,*]

[1]*Academy of Aviation Services and Music, Nanchang Aviation University, Nanchang 330013, Jiangxi, China*

[2]*School of Computer Information and Engineering, Jiangxi Normal University, Nanchang 330000, Jiangxi, China*

[3]*School of Education, Jiangxi Institute of Applied Science and Technology, Nanchang 330000, Jiangxi, China*

[4]*Institute of Music, UCSI University, Kuala Lumpur, Malaysia*

**Corresponding Author.*

Abstract:

Computer music automatic composition not only allows the general public to experience the joy of music creation, but also provides creative inspiration for professionals, greatly reducing the time for music creation. However, there is relatively little research on the use of technological means to create Guzheng music. This paper will conduct research from a new perspective by introducing joint cross attention and automating composition. The first step is to collect Guzheng music data, the second step is to process Guzheng music data, the third step is to extract Guzheng music features, the fourth step is to combine cross attention data processing, and the last step is to automatically predict the work. Through experimental data analysis, it was found that the model quickly converged and reached its peak, meeting the requirements for model migration implementation. After migration, the model can automatically compose music, meeting the needs of ordinary users.

Keywords: Guzheng, joint cross attention, lstm, transformer, automatic composition, Safety technology

INTRODUCTION

With the development of machine learning technology and the advent of the big data era, more and more work is beginning to combine artistic creation with computer technology, becoming a new research direction in the field of artificial intelligence. Music composition refers to the process in which composers create music with certain acoustic aesthetics and complex emotions according to rhythmic rules, requiring strong creativity, which is a unique human ability. From the initial inspiration of music to the final presentation of sound, it requires three processes: sheet music composition, instrument arrangement, and performance [1]. The most basic of these is the process of sheet music creation, which generally refers to the aspect of composition. In modern music composition, there are various methods such as (main) melody first, harmony first, rhythm first, etc., which make the composition ideas increasingly diversified.

Automatic composition has always been a topic of discussion and research in the music and mathematics industries, dating back to medieval musicians designing a rule-based vowel to pitch mapping method to generate music. The method of automatically generating music using dice became popular in 18th century Europe, with famous musicians such as Haydn, Bach, Mozart, and others joining in for research and experimentation [2]. With the advancement of technology and the rapid development of the computer industry, the topic of computer automatic composition has gradually attracted people's attention and research. The implementation of some theories and methods from past composition and automatic composition on computers is collectively referred to as computer algorithmic composition. Computer algorithm composition is usually based on certain rules and probabilities, and commonly used models or methods for computer algorithm composition include Markov models, generative syntax, cellular automata, genetic algorithms, transition networks, etc. [3,4]. However, computer algorithm composition lacks generalization ability, and these models must manually define music based on music rules to generate music under corresponding rules. The definitions and rules that exist in different types of music are not the same, and no definition or rule can fully cover all types of music. After the rules are determined, these models can only generate a fixed type of music under that rule. Hadjeres et al. demonstrate its efficiency on the task of generating melodies satisfying unary constraints in the style of the soprano parts of the J.S. Bach chorale harmonizations [5]. Chen et al. expound the representation and coding methods of symbolic music, and focuses on the induction, comparison and analysis of deep learning-based models, which are divided into three categories according to different basic structures [6]. Nan et al. proposes a new model to explain why

melody variations of tonal music observe the power law, which is quite universal for many natural and artificial systems [7].

Since the founding of the People's Republic of China, due to social stability and prosperous economic and cultural development, ethnic instrumental music has received attention, and more and more people have come to study and research the art and culture of Guzheng music. Especially in the 21st century, research on various aspects of Guzheng music is becoming more mature and shining brightly. However, there is still relatively little analysis and research on the creation of Guzheng music using composition theory techniques and modern information technology methods. The polyphonic structure derived from the monophonic melody in traditional Chinese music is not reflected in a simple, pulsating external form for Western polyphony, nor is it done by chance. It is based on traditional Chinese polyphonic thinking and creates a polyphonic structure with distinct visual contrasts. This vividly contrasting three-dimensional melody structure has a wide range of applications in the creation of traditional Chinese Guzheng music.

This paper conducts research from a new perspective, studying the Mel spectrogram features and simplified spectrogram features of Guzheng music. In order to calculate the joint features between them, a joint cross attention method is proposed to calculate the cross features between the Mel spectrogram features and simplified spectrogram features of Guzheng music, completing the automated creation of Guzheng works. Pre train the model using the MAESTRO Dataset dataset [8], and then train it with 10 different types of Guzheng music works to obtain a transferable model that can automatically compose music and meet the needs of ordinary users.

RELATED KNOWLEDGE

Fundamentals of Music Theory

Music, broadly speaking, is the arrangement of sounds in a certain way, and is an art of time. And sound is a physical phenomenon generated by the vibration of objects. The sound in music can usually be divided into musical notes and noise. Music refers to the constantly changing and regularly vibrating sound, which sounds pleasant to the ear. The sound produced by instruments such as the Guzheng is music. And sounds with unclear pitch changes and irregular vibrations are called noise. Both music and noise have four major characteristics: pitch, length, intensity, and timbre. The frequency of vibration determines the pitch, the duration of vibration determines the length of the sound, the amplitude of vibration determines the intensity of the sound, and timbre is determined by various factors such as the mode of vibration and overtones. The pitch in music generally refers to the pitch of musical notes. The pitch of a musical sound is a set composed of a finite number of levels, which is an auditory attribute for humans. Depending on this attribute, the sound can be arranged from low to high, and as the pitch changes, it ultimately forms a melody. What kind of pitch does each element in the set correspond to, and what is the vibration ratio of each adjacent sound, are issues that need to be considered in rhythm. The currently recognized and widely used legal systems in the world include the law of five degrees of mutual growth, the law of purity, the law of the mean, and the law of twelve averages. The most commonly used rhythm system among them is the twelve tone equal temperament, and now most music is created according to this rhythm system [9].

LSTM

LSTM (Long Short Term Memory) [10] is a variant of recurrent neural networks and is also the spectral feature extraction method constructed in the present invention. Long short-term memory networks not only inherit the advantages of traditional recurrent neural networks, but also overcome the problem of gradient explosion or disappearance in recurrent neural networks. They can effectively process sequence data of any length and capture long-term dependencies of data. The total amount of data processed and training speed have been greatly improved compared to traditional machine learning models. Long short-term memory networks are composed of many repeating units, which are called memory blocks. Each memory block contains three gates and one memory unit, with the three gates being the forget gate, input gate, and output gate.

Forgetting Gate: The first internal structure of LSTM is the forgetting gate, which determines which output information from the previous position should be forgotten. Mainly for sigmoid structure, because the output range of sigmoid is 0~1. The input of this layer is connected to the output of the previous position and the input

of the current position, and then undergoes linear transformation. Finally, the sigmoid activation function is used to select which information is forgotten. Formula 1 is as follows:

$$f_t = \sigma(W_f \cdot [h_{t-1}, S_t] + b_f) \tag{1}$$

Among them, f_t represents the output feature, b_f represents the bias term, W_f represents the weight parameter matrix, and $\sigma(\square)$ represents the sigmoid activation function.

Input gate: The second internal structure of LSTM is the input gate. It mainly consists of two parts. The input is also connected by the output h_{t-1} of the previous position and the input S_t of the current position, and then undergoes linear transformation. The first part updates the information, and the second part determines which parts need to be updated. Firstly, the sigmoid layer determines which values need to be updated, and secondly, the layer updates the input information. The entire calculation process is shown in formula 2 and formula 3.

$$i_t = \sigma(W_i \cdot [h_{t-1}, S_t] + b_i) \tag{2}$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, S_t] + b_C) \tag{3}$$

Among them, i_t is the information that needs to be updated, W_i is the weight parameter matrix, b_i is the bias term, $\tilde{C}_t$ is the output state of updating the previous position, and is the weight parameter matrix.

Output gate: The last structure inside LSTM is the output gate, which first determines which information needs to be output through the sigmoid layer, then updates the neuron state information by multiplying the neuron state, and finally multiplies to obtain the desired output part.

Transformer Model

A model proposed by Google [11] that applies Attention to the encoder decoder structure is presented in this paper, which presents a novel Transformer model. The overall architecture of Transformer can be divided into three parts: 1) Input part; 2) Output section; 3) Encoder section. The encoder part is composed of N stacked encoder layers, each consisting of two sub layer connection structures. The first sub layer connection structure includes a multi head self-attention sub layer, a normalization layer, and a residual connection. The second sub layer connection structure includes a feedforward fully connected sub layer, a normalization layer, and a residual connection.

In order for the model to capture features from different subspaces, the multi head attention mechanism applies attention to multiple heads, as shown in formula 4.

$$MultiHead(Q,K,V) = [head_1, head_2, ..., head_n]W^0 \tag{4}$$

The three specified inputs Q (query), K (key), and V (value) are represented as query vectors, key vectors, and value vectors. Among them, $head_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$ the attention mechanism is shown in formula 5.

$$Attention(Q,K,V) = \text{soft}\max(QK^T / \sqrt{d_k})V \tag{5}$$

Then, the calculation result of multi head attention is obtained through a formula, which represents the query under the action of key and value. The feedforward fully connected layer in the encoder is used to consider that the attention mechanism may not fit the complex process well enough, and to enhance the model's ability by adding it. The calculation process is shown in formula 6.

$$FFN(x) = \max(0, xW_1 + b_1)W_2 + b_2 \tag{6}$$

Among them, $\max(0,...)$ is the ReLU activation function.

In the process of inputting to each sub layer and normalization layer, residual connections (skip connections) are also used, and this part of the structure is called sub layer connections as a whole. In each encoder layer, there are two sub layers, and these two sub layers, together with the surrounding linking structure, form two sub layer connection structures.

The linear layer obtains the output of a specified dimension by linearly changing the result of the previous step, which is the function of transforming dimensions. The softmax layer scales the numbers in the last one-dimensional vector to a probability range of 0-1, satisfying their sum to be 1, and combines them with the music emotion type to obtain the prediction result.

RESEARCH ON CHINESE GUZHENG AUTOMATIC COMPOSITION BASED ON JOINT CROSS ATTENTION

This paper mainly studies the automatic composition of Guzheng, including the first step of Guzheng music data collection, the second step of Guzheng music data processing, the third step of Guzheng music feature extraction, the fourth step of joint cross attention data processing, and the fifth step of automatic prediction of works. As shown in Figure 1.

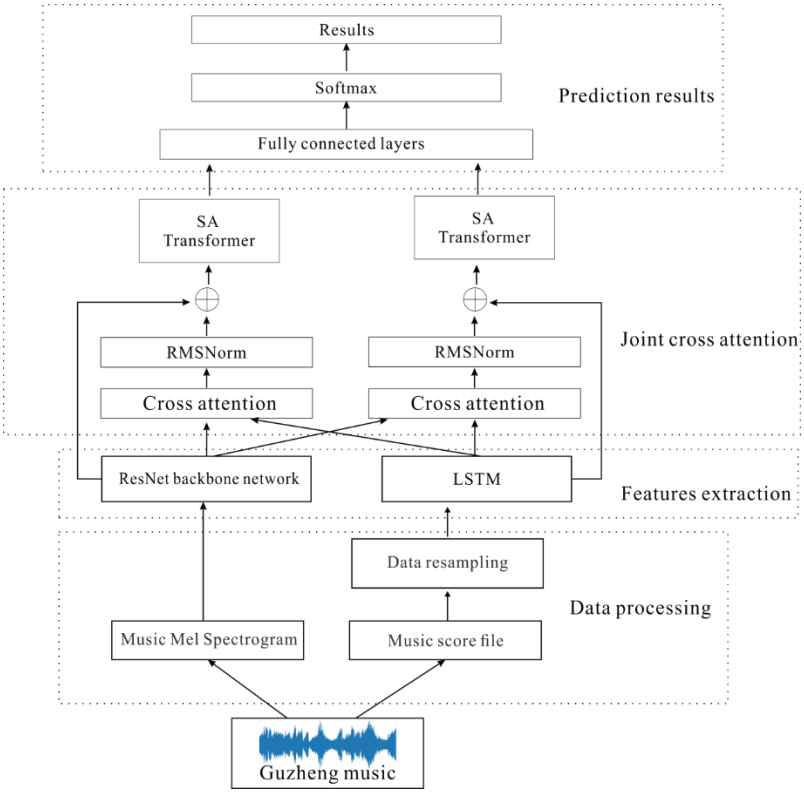

Figure 1. Framework of Chinese Guzheng automatic composition based on joint cross attention

Music Data Collection

According to the aesthetic characteristics of the audience, music is also constrained and influenced by factors such as pitch and scale. Through the study of Guzheng music, this paper collected 10 styles of Guzheng music, and the differences between each style can also be felt. This paper also used the MAESTRO Dataset for pre training the model. MAESTRO is a dataset consisting of approximately 200 hours of performance, finely arranged between note labels and audio waveforms (approximately 3ms). This dataset contains approximately 200 hours of paired audio and MIDI recordings from ten years of international piano competitions. MIDI data includes keystroke speed and the position of sustain/slow/bass pedals. The audio and MIDI files are aligned with a precision of 3 milliseconds and cut into individual music pieces with annotations for the composer, title, and performance year. The uncompressed audio is of CD or higher quality (44.1-48 kHz 16 bit PCM stereo).

Data Processing

Music mel spectrograms

The use of music Mel spectrograms is more in line with the auditory characteristics of the human ear. According to physiological research, it has been found that the human ear has varying degrees of auditory responsiveness to sound signals of different frequencies. The perceptual effect of the human auditory system on sound signals shows a logarithmic relationship with the frequency of the sound signal, that is, the relationship between the Mel frequency B (f) and the actual frequency f can be expressed by the following formula 7.

$$B(f) = 2595\lg(1 + f / 700) \tag{7}$$

Resampling of Guzheng music simplified score

The quality of music composition is closely related to the representation of music features. The audio in the dataset cannot be directly used for model training. Before data training, a series of processing is required to make the data meet the input format requirements of the model. Due to the different lengths of music files, audio files need to be processed. In this paper, the minimum processing time for simplified music files is 0.25 seconds, defined as a single character block. Every 32 character blocks form a sample, and each piece of music can be processed as shown in formula 8.

$$S = \{S_i \mid S_i = left(M, i, 32), 0 < i < l\} \tag{8}$$

Among them, S is the sample set of music M, representing the i th sample, and *left* function refers to taking a string of length 32 from the i th position as the i th sample.

Feature Extraction

Using Long Short Term Memory (LSTM) networks to extract features from sampled samples and obtain music score features. Using ResNet backbone network to extract features from music Mel spectrogram and obtain music Mel spectrogram features.

Joint Cross Attention

This paper proposes a joint cross attention Transformer encoder for transfer learning in piano music emotion recognition. In order to cross the information between the music Mel spectrogram feature encoder and the music simplified spectrogram feature encoder, joint cross attention was constructed. The music Mel spectrogram feature encoder will adjust the attention of the output of the music simplified spectrogram feature encoder to obtain the music simplified spectrogram feature encoder information related to the encoding position of the music Mel spectrogram feature. Similarly, the music Mel spectrogram feature encoder information related to the encoding position of the music simplified spectrogram feature can also be obtained. It can enable the encoder to effectively model the context of the current generated location.

The structure of the music Mel spectrogram feature encoder is the same as that of the music simplified spectrogram feature encoder. All are composed of sub layers connected in series, including cross self-attention mechanism, root mean square normalization layer, and residual connection layer.

In order to capture the features of two different subspaces, the music Mel spectrogram features S_1 are used as the key vector K_1 and value vector V_1 of attention, and the music simplified spectrogram features S_2 are used as the query vector Q_2 of attention. Calculate the attention weight matrix using the query vector Q_2 and key vector K_1, and multiply it with the value vector V_1 to obtain the music Mel spectrogram encoder information as shown in formula 9.

$$attention(Q_2, K_1, V_1) = softmax(\frac{(W_{Q_2} S_2)(W_{K_1} S_1)^T}{\sqrt{d}}) W_{V_1} S_1 \tag{9}$$

Among them, W_{Q_2} is the parameter matrix of the query vector Q_2, W_{K_1} is the parameter matrix of the key vector K_1, W_{V_1} is the parameter matrix of the value vector V_1, and d is the data dimension.

Similarly, the encoder information for music score features can be obtained as shown in formula 10.

$$attention(Q_1, K_2, V_2) = softmax(\frac{(W_{Q_1} S_1)(W_{K_2} S_2)^T}{\sqrt{d}})W_{V_2} S_2 \quad (10)$$

As the number of network layers increases, the parameters may start to become too large or too small after calculation through multiple layers, which may cause anomalies in the learning process and slow convergence of the model. The root mean square normalization introduced in this paper abandons the centralization operation [12]. The normalization process only implements scaling, and the scaling factor is root mean square. When the mean is 0, it is root mean square normalization. Root mean square only performs scaling and does not change the original distribution of the data, which is beneficial for the stability of the activation function output (scaling does not change the vector direction). The calculation process is as follows:

$$a_i = \sum_{j=1}^{m} w_{ij} attention_j, \overline{a_i} = \frac{a_i}{RMS(a)} g_i, RMS(a) = \sqrt{\frac{1}{n}\sum_{i=1}^{n} a_i^2}, \ y_i = f(\overline{a_i} + b_i) \quad (11)$$

Among them, w_{ij} represents the weight of the neuron, $RMS(a)$ represents the a_i root mean square value, a_i represents the weighted calculation result, $f(.)$ represents the non-linear function of element calculation, y_i represents the output result, g_i represents the gain parameter used to readjust the standardized total input, $\overline{a_i}$ represents the result a_i after root mean square normalization, and n represents the number of a_i.

Prediction Results

Using a linear layer to linearly transform the extracted results to obtain the output of a specified dimension, this is the function of transforming dimensions. The softmax layer scales the numbers in the vectors of the linear fully connected layer to a probability range of 0-1, ensuring that their sum is 1, and obtains the predicted results.

EXPERIMENTAL ANALYSES

Guzheng Music Data Sample

Text training uses Guzheng music, with a sampling rate of 44100Hz and a bit rate of 128kBit/s. Taking the work "Evening Sunshine" as an example, its lyrical and romantic music style is the main tone, which brings a warm and gentle feeling to the audience and easily resonates emotionally. In today's fast-paced social life, such lyrical music can soothe and relax the hearts of listeners. Its melodic characteristics are rich in layers, the transition between liveliness and gentleness, and the ease of switching between gentleness and liveliness. On the one hand, the gentle melody can express a peaceful and beautiful atmosphere; On the other hand, the dynamic parts add vitality and interest to the music, avoiding monotony and mediocrity. This contrast and variation make the music more vivid and expressive. The audio waveform is shown in Figure 2.

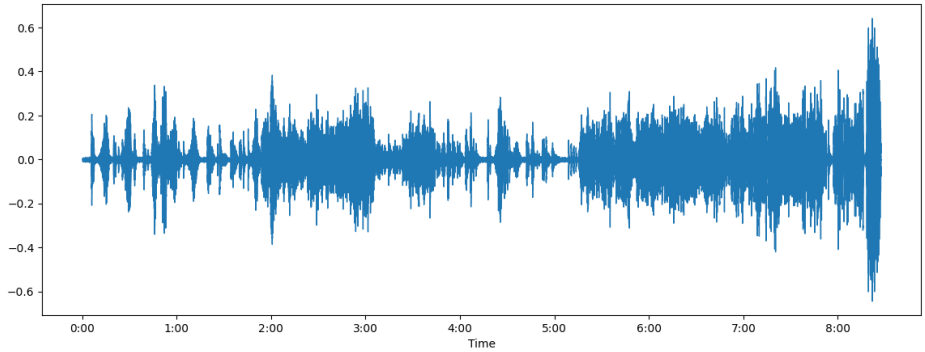

Figure 2. Audio waveform of "evening sunshine"

Guzheng Music Mel Spectrogram

The use of music Mel spectrograms is more in line with the auditory characteristics of the human ear. According to physiological research, it has been found that the human ear has varying degrees of auditory responsiveness to sound signals of different frequencies. The music chart of "Evening Sunshine" is shown in Figure 3.

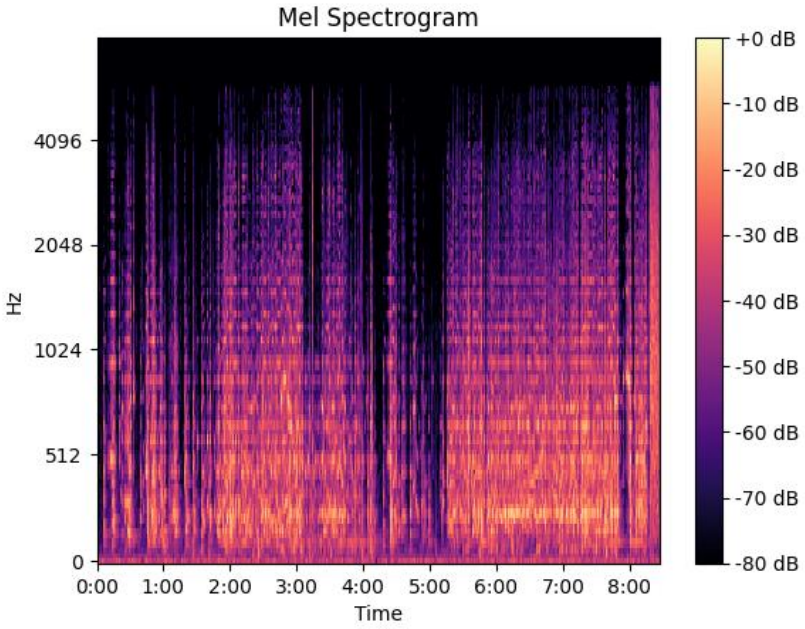

Figure 3. Music spectrogram of "Evening Sunshine"

The unit of frequency is HZ, and the frequency range that the human ear can hear is 20-20000HZ. However, the human ear is not linearly sensitive to HZ units, but is sensitive to low HZ and insensitive to high HZ. Converting HZ frequency to Mel frequency results in a linear perception of frequency by the human ear.

Guzheng Music Simplified Score Analysis

By extracting the simplified score of the music "Evening Sunshine ", the content of the simplified score can be obtained as shown in Figure 4.

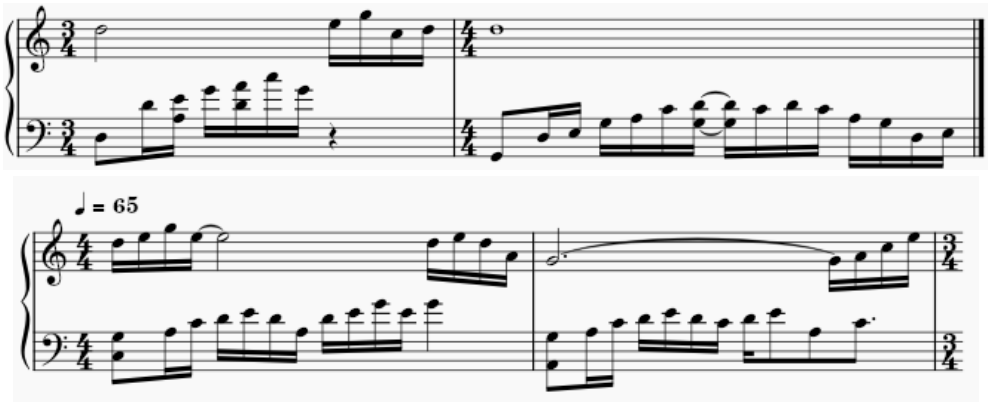

Figure 4. Content of the simplified score of "Evening Sunshine"

From the Figure 4, it can be seen that the melody is rich in layers and can easily transition between gentleness and liveliness. On the one hand, the gentle melody can express a peaceful and beautiful atmosphere; On the other hand, the dynamic parts add vitality and interest to the music, avoiding monotony and mediocrity.

Training Experiment Data Analysis

From the perspective of accuracy and speed, performance is crucial for making it suitable for practical applications in real life. In the Figure 5 and Figure 6, it can be seen that the accuracy of the model increases with the increase of epochs/iterations, while the loss of the model decreases with the increase of epochs/iterations. If the training results do not meet the standard, it indicates that some algorithms are incorrect. This may be a case of underfitting or overfitting, which can be corrected by modifying layers/parameters or improving the dataset.

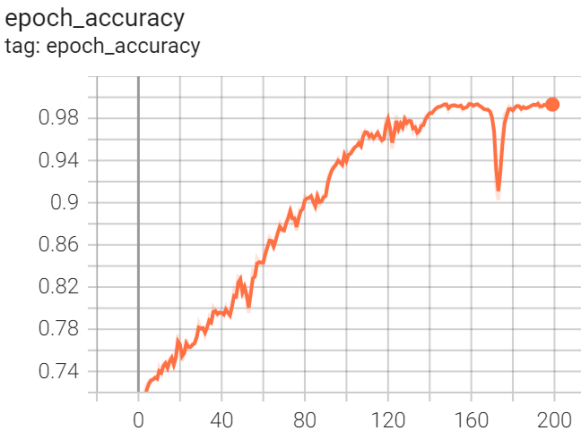

Figure 5. Epoch accuracy diagram

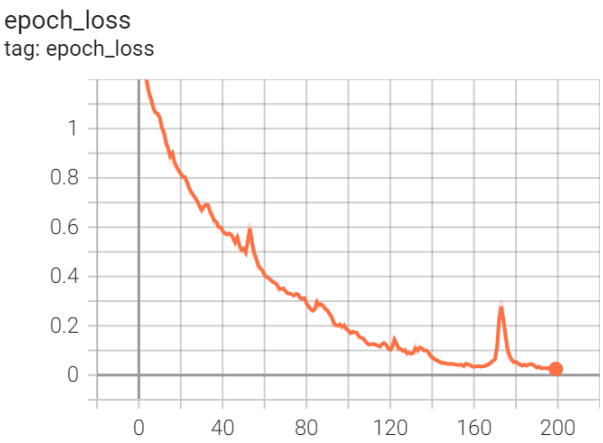

Figure 6. Epoch loss diagram

During the training process of Guzheng music samples, it was found that at epochs=159, the accuracy reached its maximum value and the loss reached its minimum value. However, the subsequent training did not improve the accuracy, but instead showed a concave shape, which should be an overfitting phenomenon. The final accuracy of 0.9935 basically meets the requirements for model transfer implementation. During the experiment, it was found that the model quickly converged and reached its peak, largely due to its good handling of the dataset, filtering out most of the noise and enabling the model to converge quickly.

Example of Automatic Composition

According to different guiding tones, the predicted simplified scores are shown in the following figure 7-9.

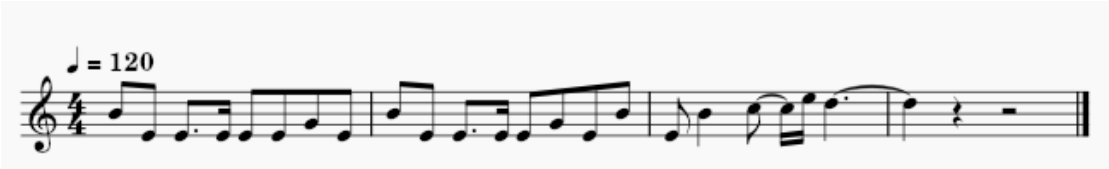

Figure 7. Simplified score 1

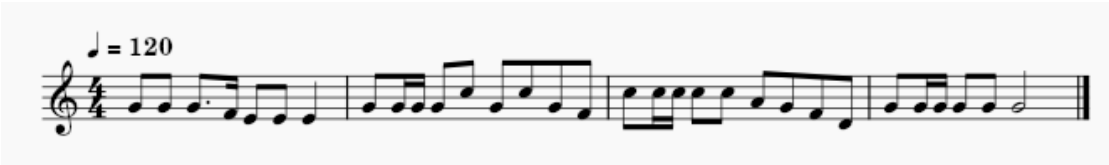

Figure 8. Simplified score 2

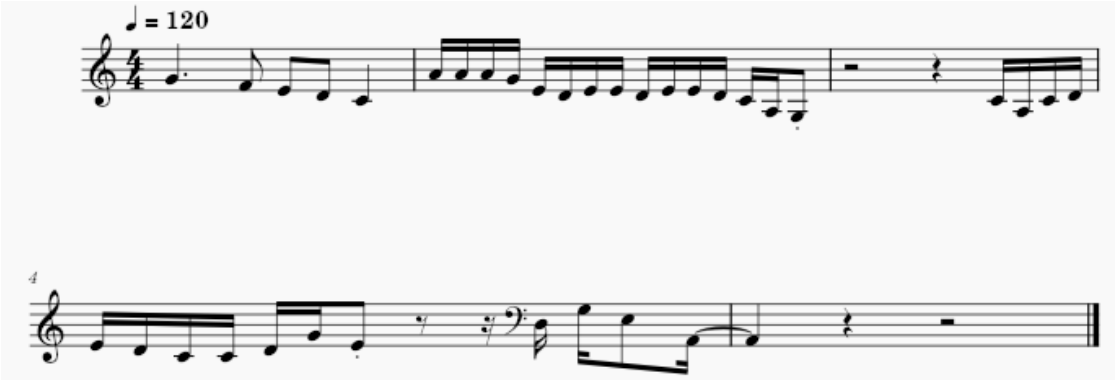

Figure 9. Simplified score 3

The above three music simplified score with a speed of 120 is a fast playing speed, and all three melodies express a cheerful mood. The rhythm uses split notes and dotted notes, with a bright rhythm and lively melody. The first simplified score is composed of rapidly repeating notes of the same pitch multiple times. The extensive use of 5 (SOL) and 1 (DOL) makes the melody jump, lively, and lively. The second point is to change the melodic style of traditional Guzheng music. The 4 (FA) and 7 (XI) notes of traditional Guzheng music are mostly transitional notes, and the 7 (XI) note is more commonly used in the Chinese melodic style dominated by pentatonic modes, which is refreshing and incorporates modern elements into traditional music, achieving a combination of tradition and modernity. The third rhythm has diverse forms, such as large cuts, staccato, and rapid alternation of sixteenth notes, all of which make the melody dynamic and impactful to the soul. The use of four types of rest symbols breaks the rhythmic pattern of strong and weak melodies, making music more irregular in rhythm.

SUMMARY AND OUTLOOK

In summary, the main purpose of this paper is to study and analyze the automatic composition of Guzheng music, accurately predict the simplified score of the entire composition based on the guiding tone, and provide assistance for the theoretical development of Guzheng music. This paper mainly completes the following aspects:

1) Through the study of Guzheng music, this paper collected 10 styles of Guzheng music, and the differences between each style can also be felt.

2) To provide creative inspiration for professionals and greatly shorten the time of music creation, this paper will conduct research from a new perspective by introducing joint cross attention to complete automated composition.

Although the model constructed in this paper has shown good performance in the automated composition of Guzheng music, there are still some shortcomings. The future prospects for improving automated composition are as follows:

The works generated by the automated creation of the model proposed in this paper are prone to the problem of style homogenization. Because it imitates and repeats existing music styles and patterns, the resulting music lacks uniqueness and personalization. If we add feature extraction of music uniqueness and personalized elements to the model, and supplement it with a large number of training samples, we should be able to solve the problem of melody similarity and rhythm similarity, making the music market more diverse.

REFRENCES

[1] Oore S, Simon I, Dieleman S, et al. This time with feeling: Learning expressive musical performance. Neural Computing and Applications, 2020, 32: 955-967.
[2] Wang Jie. Music and Mathematics. Beijing: Peking University Press, 2019.
[3] Jing Yinji, LI Shengchen. Computer Automated Composition Survey: A Generic Model. Journal of Fudan University (Natural Science), 2020, 59(06): 639-657.
[4] Hernandez-Olivan C, Beltran J R. Music composition with deep learning: A review. Advances in Speech and Music Technology: Computational Aspects and Applications, 2022: 25-50.
[5] Hadjeres G, Nielsen F. Anticipation-RNN: enforcing unary constraints in sequence generation, with application to interactive music generation. Neural Computing and Applications, 2020, 32: 995-1005.
[6] Chen Jishang, Abudukelimu Halidanmu et al. Review of Application of Deep Learning in Symbolic Music Generation. Computer Engineering and Applications, 2023, 59(09): 27-45.
[7] Nan N, Guan X, Wang Y, et al. Common quantitative characteristics of music melodies-pursuing the constrained entropy maximization casually in composition. Science China (Information Sciences), 2022, 65 (07): 240-242.
[8] Hawthorne C, Stasyuk A, Roberts A, et al. Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset//International Conference on Learning Representations, 2019.
[9] Han Baoqiang. An Introduction of Musical Acoustics. Beijing: People's Music Publishing House, 2016.
[10] Hochreiter S, Schmidhuber J. Long Short-Term Memory. Neural Computation, 1997, 9(8): 1735-1780.DOI:10.1162/neco.1997.9.8.1735.
[11] Ashish Vaswani, Noam Shazeer, Niki Parmar et al. Attention is all you need//In Proc. of the 31st International Conference on Neural Information Processing Systems, USA, 2017: 5998–6008.
[12] Biao Zhang, Rico Sennrich. Root Mean Square Layer Normalization// In Proc. of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada, 2019.