Deep Learning in Computer Aided Pop Music Creation System

Yong Hu[1,*] and Xi Fang[2]

[1]School of Music, Hebei Institute of Communication, Shijiazhuang, Hebei 050000, China, huyong@hebic.edu.cn
[2]Bohong Culture Development Beijing Co., Lt, Beijing, Beijing 100097, China, fangxi5617@163.com

Corresponding author: Yong Hu, huyong@hebic.edu.cn

Abstract. Music has its own style, and it is necessary to control the style of music. However, there is not enough data to train the corresponding music generation model in reality. This paper studies the application of DL(Deep learning) in computer-aided pop music creation system. Based on multi-task learning and a network structure of pop music style recognition and generation are proposed. In a musical sequence, the correlation between each step before and after is very strong. Therefore, the nerve cells in the hidden layer are designed to be relatively complex, and special components such as doors are added. The value of each unit in the output layer is a probability value between 0 and 1, and the note category with the highest probability value is selected as the final actual predicted note output. The network can complete the task of note prediction and generation well, and the generated music has high quality, which meets the expected requirements of this paper.

Keywords: Deep Learning; Computer-Aided; Pop Music; Neural Network.
DOI: https://doi.org/10.14733/cadaps.2023.S7.142-152

1 INTRODUCTION

AI (artificial intelligence) has made remarkable achievements in traditional fields, and even surpassed the human level in some fields. AI is gradually entering the art field. At present, many research institutions are promoting this process [1]. The relationship between human beings and computers is getting closer and closer, Intelligent computer accompaniment is a new direction in AI field, and music score following is an aspect of intelligent computer accompaniment [2]. Creation is the core of music production and the beginning of the whole music production activities. Only after the creation and production of music products will there be other production activities such as singing and playing production, music dissemination and music consumption. The final form of music creation is usually music score [3]. Composers condense their thoughts and feelings, feelings about life, etc. in the music score. Singers convey the composer's expression intention by reading and understanding the music score information with their voices or musical

instruments. It is just like music score, except that after the score is finished, the performer performs the music, and the music sequence of sequencer is handed over to the synthesizer for music performance. Then, it has an early budding intelligent mode by means of music playing with the instruments tied together [4].

The sound operation of social music production is an important aspect of social harmony and an important component of China's socialist spiritual civilization construction. Therefore, studying the relationship between digital technology and music production is helpful to fully understand the application of digital technology in pop music production. The repetition, tone sandhi, interval, rhythm, arrangement and combination of pitch in music, parallelism in musical form, etc. can all be described and modeled by algorithms [5]. Schiavoni et al. [6] think computer music production software is becoming more and more popular, greatly enriching the form of music creation. In the process of music creation, it is necessary to create music scores through computer music production software. It uses computer assisted virtual keyboard, so you need something to turn your computer keyboard into a virtual MIDI keyboard. Help professional musicians and non-professional music lovers quickly input their imagined notes into music scores. The rhythm and beat in the computer-assisted music score, the speed of the music and the strength of the notes can be accurately controlled. Schumacher and Wanderley [7] believed that music lovers can not only create timbres through user settings, but also program images, graphics, etc. This creative mode of composition assisted by Max/Msp has great contingency and randomness. In various forms of electronic music competitions and academic conferences, most of the creators' electronic music works are created through the Max/Msp platform. These music works assisted by computer music production software have gradually increased over time. Computer aided programming is mainly for audio, which has many different audio modules. Max/Msp, a computer music production software, can be applied to music creation, mainly in algorithmic composition, electronic music composition and sound design. Stoller et al. [8] evaluated the previous work and improved its limitations, especially ignoring the loudness change during shearing and the non-intuitive control of the output music structure. By using a * algorithm to greatly speed up the search for the best output trajectory, our interface responds faster. Listening experiments have proved the improvement of perceived audio quality. Xu and Zhao [9] In music creation, the development of timbre synthesis technology is very fast, which fully shows that the integration of music synthesizer and computer technology in China is very close. This shows that the integration of technology and art has become an inevitable development trend. Technology can become an important carrier of art communication, and it can also shine more brightly because of the charm of art. It can provide more inspiration for the majority of music creators. Therefore, the integration of advanced computer technology into computer music can release the essence of music art. It can be converted into signals that can be received by headphones or speakers, and corresponding graphics, music scores, icons, etc. can also be displayed on the software. With the constant change of music, graphics will also change. At the same time, you can switch between multiple interfaces. In this way, the music can be modified in time to continuously improve the quality of music. When using computer music production software, the software can automatically convert the received playing information in time. Yang and Li [10] At present, only a few music compositions are likely to show the effect of the works in the form of bands. Today, the powerful interactive function of computer music that integrates text, graphics, audio and video is used. This not only reduces the time for teachers to copy music scores temporarily in class, but also improves the classroom efficiency. More importantly, the students felt the rich sound effects brought by the boring theoretical explanation and the thin piano color before. It is difficult for most students to have the opportunity to show their works in a real way so as to constantly modify and improve them. The initiative of students' hearing, vision, feeling and thinking processing is fully mobilized to obtain more information and deepen students' understanding and memory of the teaching content. Have a unified understanding of the knowledge concepts learned from words to sound, and achieve the combination of theory and practice in a real sense. A method for detecting performance error is developed, and then the local area around the alignment error is realigned by

HMM. The experimental results show that this method has short calculation time and high accuracy.

2 RESEARCH METHOD

2.1 Music Domain Knowledge

The production of popular music has evolved from the analog era to the digital era, and each link in the music production has a sequence in the digital process. Therefore, an exact time node cannot be given as the dividing line of the analog-digital era. When we compare the production efficiency of pop music in the era of modulus, we don't focus on the whole production and operation system, but by analogy with the local comparison of production systems, we illustrate the influence of analog technology and digital technology on the production of pop music from point to point. The core component of the recording system is the recording equipment, and the biggest difference between analog recording and digital recording system lies in this. In the specific production, when the quality identity is to be maintained, the quantity will inevitably be reduced, so that the production efficiency will be reduced. Through the comparative analysis of analog recording mode and digital recording mode in recording production, it is more efficient than that under analog conditions.

On the basis of DL machine learning, it has been further improved. Because DL architecture contains more layers of networks, more features can be obtained when analyzing features, which improves the learning ability of the network. At present, DL has been used in various fields, such as finance, security, manufacturing and so on. Neuron can also be called perceptron, which is the most basic unit of neural network. The neural network is shown in Figure 1:

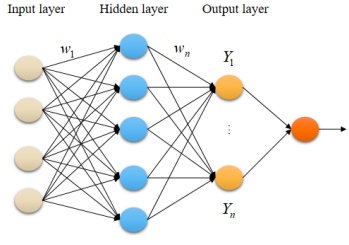

Figure 1: Neural network diagram.

If there are many hidden layers of neural network, it belongs to DL category. DL mainly studies different deep neural networks, and then uses them to solve different problems.

The input of the input layer is a vector $x_1, x_2, \cdots x_n \middle| x_i \in R$, and each input should have a weight w, while the other offset is b, and the value of the offset is generally 1, which can be recorded as w_0. The formula for calculating the sum of weights is as follows (1):

$$z = \sum_{i=1}^{n} w_i x_i + b \tag{1}$$

The output can be recorded as y, and the value of y can be calculated by activating function $g(z)$. There are many options for activating function, and the calculation formula is shown in (2):

$$y = g(z) \tag{2}$$

$$\sigma(x)=\frac{1}{1+\exp(-x)} \tag{3}$$

$$E=-\sum_{n=1}^{N}\left[r_n \ln y_n+(1-r_n)\ln(1-y_n)\right] \tag{4}$$

Cross entropy refers to the concept in information theory, which is used to measure the distance between two probability distributions.

2.2 Melody and Arrangement Generation

Music is a form of creative art. It has been marked as a symbol of a special composer, a group, a country or a culture at different times in history. People from all over the world have their own music at different times. There will inevitably be some subtle changes in the inheritance of rhythm. We assume that these small changes will not affect our research for the whole creation. When constructing music network, in order to abstract music notes into network nodes, it is necessary to adopt digital coding of music scores. If the network is found to have a short average shortest path and a high aggregation coefficient, the frequency of common notes appearing together in a tune can be explained, thus the motivation of the whole tune can be excavated and the style of the work can be reflected.

Music arranger has several instruments to play at the same time. If the harmony between the instruments can't be guaranteed, it will be chaotic, noisy and counterproductive. Therefore, it is very important to ensure the coordination of multiple musical instruments. How to learn the characteristics of multiple musical instruments in the learning process is a problem worth discussing. To solve these problems, this paper proposes a multi-instrument joint arrangement model based on multi-task learning (Figure 2).

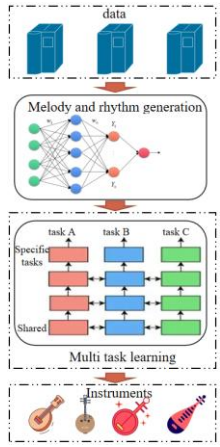

Figure 2: Melody and arrangement generation framework.

In the part of melody generation, a model based on the cross generation of chord rhythm and melody is proposed to improve the musical interval relationship and learn the structure of the segment. In the part of arrangement generation.

The previous rhythm R_{t-1} and melody M_{t-1} are multiplied by the embedding matrix E_r, E_m to represent a high-dimensional vector, and then the $\overline{R}_{t-1}, \overline{M}_{t-1}$ representation is obtained as follows:

$$\overline{R}_{t-1} = E_r R_{t-1}, \qquad E_r \in R^{V_r * d} \tag{5}$$

$$\overline{M}_{t-1} = E_m M_{t-1}, \qquad E_m \in R^{V_m * d} \tag{6}$$

V_m, V_r is the vocabulary size of notes and durations.

In this paper, the notes generated by the Actor network at the current moment are regarded as the selected action a_t, where $a_t \in A$, A are the set of actions, and the sum of the notes generated at the previous moment is regarded as the current state $s_{1:t-1}$:

$$\left(a_1, a_2, \cdots, a_{t-1}\right) \to s_{1:t-1}, \qquad s \in S \tag{7}$$

From the selected note a_t at the current time and the previous state $s_{1:t-1}$, the stat $s_{1:t}$ at the next time can be obtained.

Tick is the basic time unit in MIDI file format, but its basic duration is different in different MIDI. Therefore, we need to convert the time unit in MIDI into the time unit in real life. To realize this process, we need to calculate how many ticks there are per second. The calculation formula is as follows.

$$t = m * n \tag{8}$$

Where t represents how many ticks per second, m represents multiple frames per second, and n represents the number of ticks in a frame. m, n can all be obtained from the header block of MIDI.

There are basic elements of music, such as tone, timbre, pitch and duration. These basic features can be directly or indirectly extracted from the frequency domain features of audio, which are calculated by Fourier transform of its original time domain signal. As shown in formula (9), it is the Fourier transform expression of transforming the time domain signal to the frequency domain.

$$X_n(k) = \sum_{m=0}^{N-1} x_n(m) e^{-j\frac{2\pi m}{N}k}, \qquad k = 1,2,\cdots,N \tag{9}$$

In frequency domain, we generally use spectrogram to represent audio features, which is the energy distribution of audio signals in various frequency segments. In spectrogram, according to the energy of frequency at that moment, its color is also different. The darker the color, the greater the energy of the frequency component, and the greater the proportion of the frequency component. On the contrary, the smaller the energy at a certain moment.

2.3 Recognition of Popular Music Style Based on DL

Spectrum flatness describes the flatness of signal spectrum amplitude. The larger the amplitude fluctuation, the more uneven the spectrum is, and the smaller the fluctuation, the flatter the spectrum is. Spectral roll-off represents the frequency at which the signal energy decays to 85% of the total energy, and describes the speed at which the signal energy decays with the frequency change, which means that the music signal has short-term stationarity. Compared with the signal in the center of a certain frame, its importance is weak. Therefore, in this paper, frames are overlapped with each other during minute hand processing, and the time difference between the initial positions of two adjacent frames is called frame shift.

In a musical sequence, the correlation between each step before and after is very strong. Therefore, the nerve cells of the hidden layer are designed to be relatively complex, and special components such as gates are added.

The linear layer can activate the table value of the function, the expression of which is shown in formula (10):

$$g(z) = z = w^T x \tag{10}$$

Here, z is the sum of the weights of the input values, and x represents the input vector.

In training, the first step is forward propagation, and the input value is the time series of the processed music. Because the LSTM network is relatively complex, it includes the input gate, the output gate and the forgetting gate. The input gate is calculated according to the following formula:

$$a_l^t = \sum_{i=1}^{I} w_{il} x_i^t + \sum_{h=1}^{H} w_{il} b_h^{t-1} + \sum_{c=1}^{C} w_{cl} s_c^{t-1} \tag{11}$$

$$b_l^t = f\left(a_l^t\right) \tag{12}$$

The forgetting gate f_t controls how much information needs to be forgotten in the internal state c_{t-1} at the last moment. The calculation formula is as follows:

$$f_t = \sigma\left(W_f \cdot \left[h_{t-1}, x_t\right] + b_f\right) \tag{13}$$

The calculation formula is as follows:

$$o_t = \sigma\left(W_o \cdot \left[h_{t-1}, x_t\right] + b_o\right) \tag{14}$$

The number of notes in the note sequence data set $notes$ will be very large. Assuming that the number of different notes in $notes$ is k. The actual output note of each training of the neural network model is one of k notes, and the specific output note can be calculated by softmax function:

$$soft\max = \left(u_j\right) = \frac{\exp\left(u_j\right)}{\sum_{q=1}^{k} u_q} \tag{15}$$

The denominator of softmax function sums the activation values of all units in the output layer. The value of each unit in the output layer is a probability value between 0 and 1, and the note category with the highest probability value is selected as the final actual predicted note output. The network structure is shown in Figure 3:

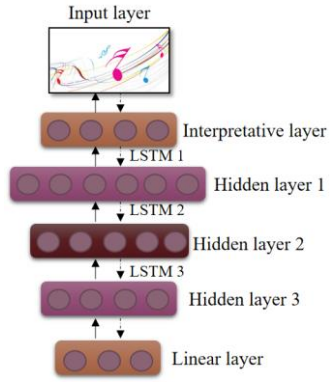

Figure 3: Identification of popular music style and generation of network structure diagram.

It can be seen that there are six layers in this network, of which the first layer is the interpretation layer, the second to fifth layers are the hidden layers of LSTM, and the sixth layer is the linear layer. Because this network is a multi-task processing network, each small genre unit contains the 2nd to 6th layer networks. The number of neurons in the input layer is 128, and other codes are very similar to this code. You only need to add codes to realize the style analysis of different genres after the explanation layer.

The neural network used in this experiment is LSTM network, and it can multitask. Because the subnet for multitask in this network is the music genre analysis subnet, the dropout layer structure is added to this subnet. The formula for adding Dropout between hidden layers is:

$$r = maskxf(Wv+b)\, r = maskxf(Wv+b) \tag{16}$$

Here, $mask$ is a binary model, which obeys Bernoulli probability distribution. When the probability value is p, the value is 1, and the other values are 0.

3 ANALYSIS AND DISCUSSION OF RESULTS

This section will analyze the experimental data set, experimental setup, experimental results of melody generation and experimental results of arranger generation in detail. In this paper, a large number of experiments are carried out on the real music data set, which includes more than 50,000 digital music files. Firstly, a large number of melody tracks and their corresponding other arranger tracks are extracted from the music, such as drums, bass, strings, guitars, etc. Then, all music tonality is converted to C major to ensure the accuracy of note representation. Finally, two bars in the music are combined into one segment as training data. Figure 4 shows the result of chord analysis of generated music.

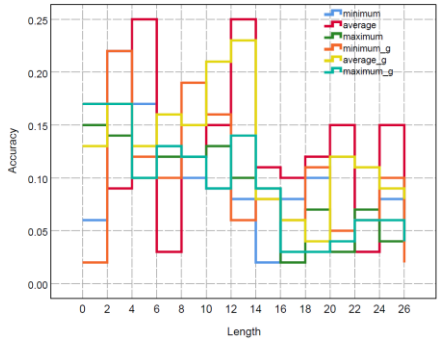

Figure 4: Rhythm distribution.

It can be seen that the generated music and the real music have similar distribution in segment length, which indicates that the model in this paper can keep good structural features.

In this section, an experiment is conducted to verify the performance of the perceptron unit model in the task of multi-track music generation. Several commonly used tracks in music are selected in the experiment: melody, drums, bass, strings and guitar. In addition to the perceptron unit model proposed in this paper, the latest multi-track generation model is selected for comparison to realize the information interaction between different tracks. As shown in Table 1 and Figure 5.

Category	*Our*	*RNN*	*HMM*

melody	0.26	0.13	0.26
drum	0.26	0.23	0.3
bass	0.35	0.25	0.35
string music	0.36	0.28	0.4
guitar	0.5	0.36	0.43

Table 1: Analysis of arrangement harmony degree.

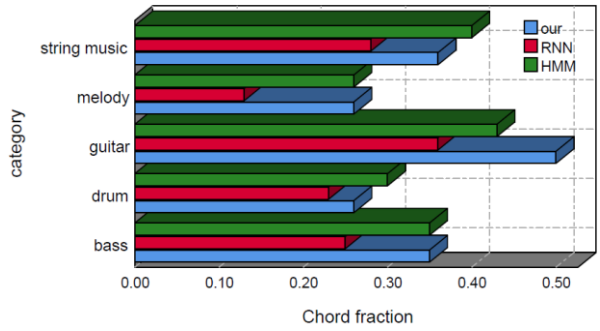

Figure 5: Analysis chart of arrangement harmony degree.

The model based on perceptron unit achieved the highest score, which was 24.416% higher than the traditional model.

In the model, the note stored on the vertex can be replaced by a new note, and the operator attached to the edge can calculate the note interval between the current note and the target note, and three target notes can be obtained from the current note, thus creating a specific selection environment. GRU (gated recurrent unit) predicts the probability of 88 notes appearing in the next step according to the notes received in the previous iteration and the newly received notes. For major and minor data sets, GRU networks corresponding to the data sets are trained and verified respectively. Figure 6 shows the learning curves of the training set and the test set, and there is no phenomenon of fitting. Table 2 lists the results on test set and training set.

		Degree of accuracy	*Cost function*
Major	Training set	0.806	0.362
	Test set	0.728	0.559
Ditty	Training set	0.902	0.302
	Test set	0.846	0.502

Table 2: GRU evaluation results.

As a result, GRU shows considerable predictability, with an accuracy of 80.6% in the training set and 84.6% in the testing set. As can be seen from the results, taking the test set as the reference target, GRU ability in minor is less than that in major. The gap between training set and test is reasonable, which reflects the generalization ability of prediction model.

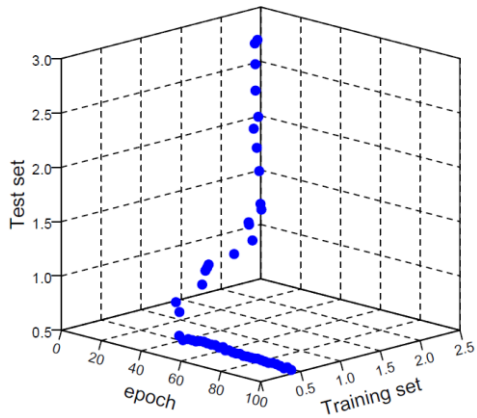

Figure 6: Learning curve of training set and testing set.

The requirements of music generation and image processing are different. Image processing focuses on the restoration and identification of samples, emphasizing "old" rather than "new", while music generation focuses on innovation, emphasizing "new" rather than "old" and emphasizing "quality" rather than "quantity". Calculate the minimum value of single-character editing (insert, replace and delete) required to change the generated music sequence into a real music sequence. This method is called editing distance. According to the rules of music plagiarism, each piece of music is set to eight bars in this experiment. The experimental results are shown in Table 3.

Model	*Melody similarity*
our	0.655
RNN	0.578
HMM	0.555

Table 3: Model melody similarity.

It can be seen that the models constructed in this paper are superior to other models in the database, and the similarity of generated music is within 0.5, and the similarity of generated music exceeds 0.5. Through this experiment, it can be concluded that the model constructed in this paper is more innovative than RNN and HMM.

Through the above operations, several LSTM-based music generation network models are finally obtained. Next, these models need to be verified. The main verification methods include: using loss curve to check the convergence of different models, and subjectively evaluating the effect of generated MIDI files, etc. Figure 7 shows the curve of loss change in the process of model training with different versions of different implementations.

Through observation, the following preliminary conclusion can be drawn: the network using bidirectional LSTM at the same time can obtain the minimum loss value in the same iteration times, that is, the idea proposed in this paper has achieved better results, so the network has better generalization ability and can better accomplish the task of note prediction and generation. The network can complete the task of note prediction and generation well, and the generated music has high quality, which meets the expected requirements of this paper.

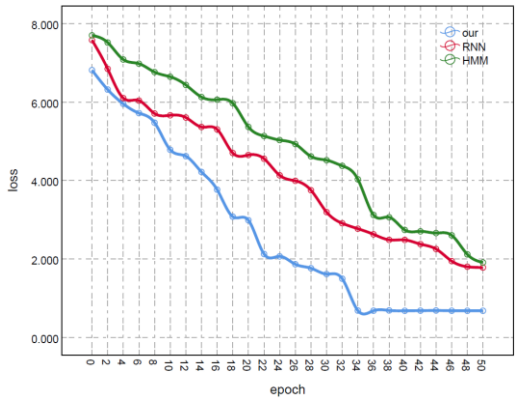

Figure 7: Experimental data of learning rate curves of different models.

4 CONCLUSION

This paper studies the application of DL in computer-aided pop music creation system. A multi-instrument joint arrangement model based on multi-task learning and a network structure of pop music style recognition and generation are proposed. In this paper, a large number of experiments are carried out on the real music data set, which includes more than 50,000 digital music files. The model based on perceptron unit has achieved the highest score, which is 24.416% higher than the traditional model, indicating that perceptron unit model.

Yong Hu, https://orcid.org/0000-0001-6392-3900
Xi Fang, https://orcid.org/0000-0002-4034-6466

REFERENCES

[1] Bülbül, B.-Ö.; Güler, M.: Can geometry achievement and geometric habits of mind be improved online? Reflections from a computer-aided intervention, Journal of Educational Technology Systems, 49(3), 2021, 376-398. https://doi.org/10.1177/00472395209652

[2] Li, Y.: Application of computer-based auto accompaniment in music education, International Journal of Emerging Technologies in Learning (iJET), 15(6), 2020, 140-151. https://doi.org/10.3991/ijet.v15i06.13333

[3] Maba, A.: Computer-aided music education and musical creativity, Journal of Human Sciences, 17(3), 2020, 822-830. https://doi.org/10.14687/jhs.v17i3.5908

[4] Onofrei, G.; Ferry, P.: Reusable learning objects: a blended learning tool in teaching computer-aided design to engineering undergraduates, International Journal of Educational Management, 34(10), 2020, 1559-1575. https://doi.org/10.1108/IJEM-12-2019-0418

[5] Quan, Y.: Development of computer aided classroom teaching system based on machine learning prediction and artificial intelligence KNN algorithm, Journal of Intelligent & Fuzzy Systems, 39(2), 2020, 1879-1890. https://doi.org/10.3233/JIFS-179959

[6] Schiavoni, F.-L.; de Faria, P.-H.; Manzolli, J.: Interaction and collaboration in computer music using computer networks: An UbiMus perspective, Journal of new music research, 48(4), 2019, 316-330. https://doi.org/10.1080/09298215.2019.1635626

[7] Schumacher, M.; Wanderley, M.-M.: Integrating gesture data in computer-aided composition: A framework for representation, processing and mapping, Journal of New Music Research, 46(1), 2017, 87-101. https://doi.org/10.1080/09298215.2016.1254662

[8] Stoller, D.; Vatolkin, I.; Müller, H.: Intuitive and efficient computer-aided music rearrangement with optimised processing of audio transitions, Journal of New Music Research, 47(5), 2018, 416-437. https://doi.org/10.1080/09298215.2018.1473448
[9] Xu, C.; Zhai, Y.: Design of a computer aided system for self-learning vocal music singing with the help of mobile streaming media technology, Computer-Aided Design and Applications, 19(S3), 2022, 119-129. https://doi.org/10.14733/cadaps.2022.s3.119-129
[10] Yang, C.; Li, Q.: Music emotion feature recognition based on Internet of things and computer-aided technology, Computer-Aided Design & Applications, 19(S6), 2021, 80-90. https://doi.org/10.14733/cadaps.2022.s6.80-90