Advances in Computer Science Research (ACSR), volume 83
8th International Conference on Mechatronics, Computer and Education Informationization (MCEI 2018)

Application of Artificial Neural Network in Chinese Folk Music

Lingyun Feng[1, a], Yulin Wu[1, b], Li Zhu[1, c] and Qian Yin[1,d *]

[1]College of Information Science and Technology, Beijing Normal University, Beijing, China

[a]201511210128@mail.bnu.edu.cn, [b]201611210907@mail.bnu.edu.cn,
[c]201511210108@mail.bnu.edu.cnl, [d]yinqian@bnu.edu.cn

Keywords: LSTM; Chinese folk music; Algorithmic composition; Interval

Abstract. In recent years, artificial neural network have been widely used in music application. Algorithm can compose music that is comparable to human performances. However, few people use algorithm to compose Chinese folk music. Applying algorithm to ethnic music composition is beneficial to its development. In this paper, we present a Chinese folk music sequence learner based on the superimposed LSTM neural network model. By considering the correlation between pentatonic scales, we create ethnic music that has more sound structures. We also conduct experiments to evaluate the quality of our music generation.

Introduction

Algorithmic composition, sometimes also referred to as “automated composition”, basically refers to "the process of using some formal process to make music with minimal human intervention" [1]. Mozer constructed CONCERT using recursive neural network[8]and trained CONCERT with backpropagation. Douglas Eck uses Long-Short Term Memory (LSTM) units to learn blues music and generate music of similar style[10]. However, artificial neural network is rarely used to generate Chinese folk music. Applying algorithm to create Chines folk music will yield much benefits for composers in their creative tasks and for customers in their entertainment, which can greatly promote the development of national music and cultural exchange at the same time.

MIDI, short for Musical Instrument Digital Interface, is one of widely used music formats. It uses the digital control signal of notes to record music. However, few music is synthesized using multiple instruments. They only focus on a single instrument or they input sheet music which lacks information of performance.[12]

Intervals play a prominent role in music generation. Melodies are formed by skipping from one pitch to another using intervals. However this factor has not been fully considered in algorithmic compositions. In this article, we first summarize the key technologies used in algorithmic composition. Then, we apply a superimposed LSTM model to generate Chinese folk music and use intervals to consider correlations between pentatonic structures. Our work is differentiated from previous works by three aspects. First, our network can learn interactions within music. The LSTM networks we use are designed to learn from specially processed MIDI data rather than sheet music which lacks information of performance. We combine instruments to achieve a fuller sound. Second, unlike the previous research, The LSTM networks is trained using different music style: Chinese folk. We apply intervals to capture stylistic constraints, which can make us generate music with more rational structure. Third, our generated music is more harmonious with a threshold set to filter mutated notes.

Related Work

At present, the main techniques used in algorithm composition are Markov model, knowledge base based on music rules, musical grammars, genetic algorithm and artificial network.

In terms of Markov model, the musical notes are selected in turn according to the conversion table, such as cybernetic composer system [2]. However, this method cannot predict subsequences of tones with more than one note. The rule-based knowledge base system can explain the choice of related

behaviors, such as the CHORAL system[3]. But it is difficult to establish knowledge guidance mechanism. Music grammar is used to compose music with statistical methods and match probability distributions of music events, such as pitch interval and rhythm such as David Cope's EMI system[4]. However, the method cannot be effectively applied to musical work without grammatical hierarchy. Using genetic algorithm to compose music mainly construct adaptive function to evaluate and select system-generated melody. More specific description can refer to [5]. For example, Biles constructed an interactive improvisation system called "GenJam" [6,7].

With advances of artificial neural network in algorithmic composition, Mozer constructed CONCERT using recursive neural network techniques[8]and trained CONCERT with backpropagation. CONCERT can create melody using note one by one, but cannot capture the global structure of music because learning gradients gradually disappear in recursive neural networks. To solve this problem, Long-Short Term Memory (LSTM) was proposed[9] and applied to music generation. Douglas Eck uses LSTM to learn blues music and generate similar style music[10]

In this paper, we use python to process the midi files, apply intervals to capture stylistic constraints and superimposed LSTM model to generate Chinese folk music. Then we carry out experiments to evaluate the generated music.

Music Generation Model

We apply the superimposed LSTM model to generate Chinese folk music. LSTMs are a special kind of Recurrent Neural Networks (RNNs), capable of learning long-term dependencies. RNNs are networks with loops in them, which can maintain information but have long-term dependency problem. All RNNs have a chain of repeating modules. In standard RNNs, this repeating module has a very simple structure, such as a single tanh layer:

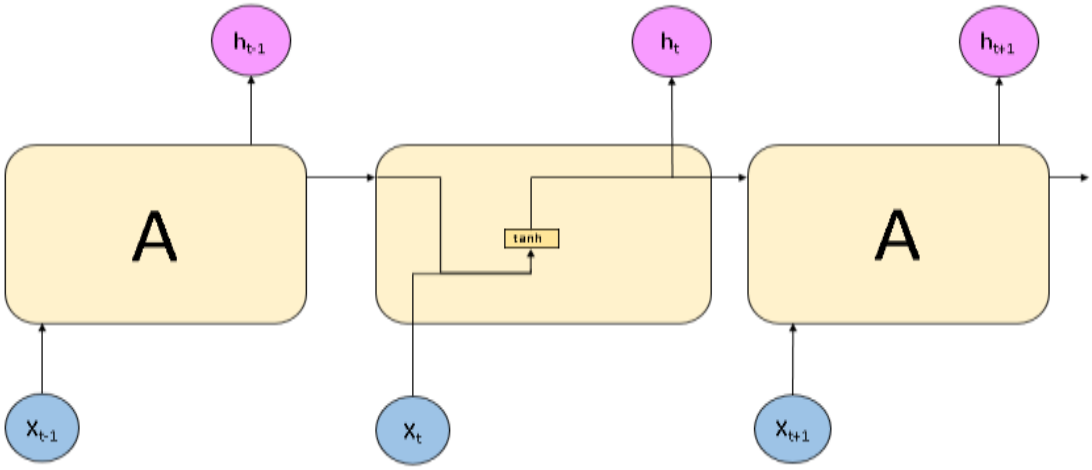


Fig 1 The repeating module in a RNN [11]

LSTMs are networks that can avoid the exploding and vanishing gradient problem when training traditional RNNs. The main idea is to store short-term memory information by structure called gates. Unlike RNNs which have a single neural network layer, there are four layers in the repeating module in LSTMs. See the specific figure below:

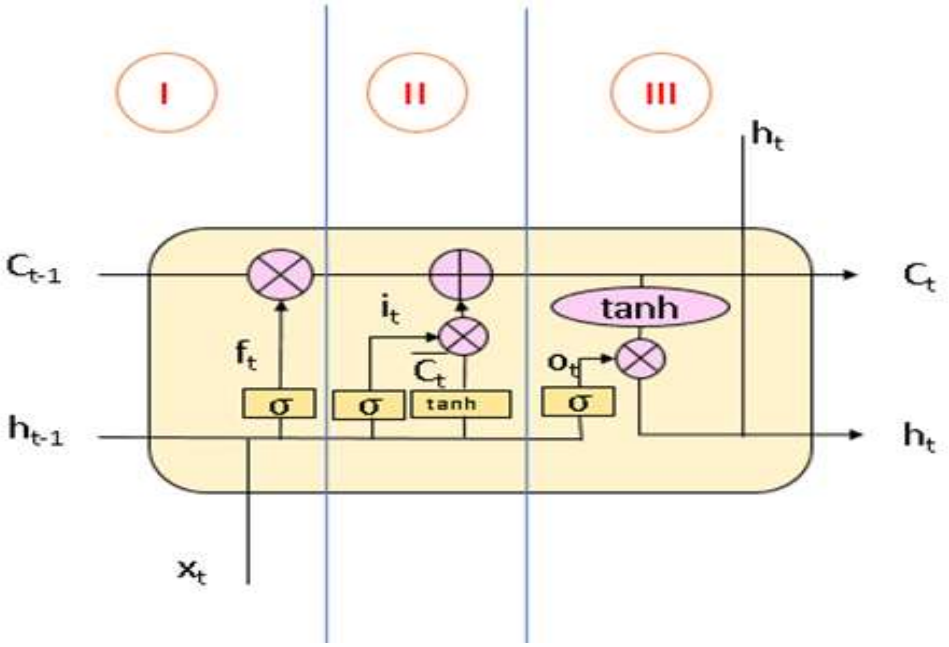


Fig 2 The repeating module in an LSTM [11]

We parse a MIDI file into measures and chords to get corpus data. Index all musical values and create a dictionary of them. Each musical value comprises a pitch and a duration. Then we construct sentence sequence and develop LSTMs for embedded vectors with indexed label sequence. For example, $x[i,t,:]$is a one-hot vector representing the value of the i-th example at time t.

Firstly we decide what information to discard by a sigmoid layer called the "forget gate layer" as shown in part I in Fig. 2. In this layer, we get sigmoid layer outputs numbers f_t in the following way:

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

x_t denotes input at current time, h_{t-1}denotes output of the last chunk. We inputh_{t-1} and x_t, output a number between 0 and 1 for each number in the cell state C_{t-1}.1 represents "completely reserve this" while 0 represents "completely discard this".

Then we decide what new information to store in the cell state. In part II, there are two layers. One is a "input gate layer" determing which value to update. The other is a tanh layer where we create a vector $\tilde{C}_t$ that can be added to the state. We get the following equation:

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

$$\tilde{\mathrm{C}}_\mathrm{t} = \tanh(\mathrm{W_c} \cdot [\mathrm{h_{t-1}}, \mathrm{x_t}] + \mathrm{b_c} \tag{3}$$

We combine both to update the old cell stateC_{t-1} into the new cell state$C_t$in the following way:

$$\mathrm{C_t} = \mathrm{f_t} * \mathrm{C_{t-1}} + \mathrm{i_t} * \tilde{\mathrm{C}}_\mathrm{t} \tag{4}$$

Then, we process the cell state through tanh to get values between -1 and 1 and multiply it by the output of the sigmoid gate:

$$\mathrm{o_t} = \sigma(\mathrm{W_o}[\mathrm{h_{t-1}}, \mathrm{x_t}] + \mathrm{b_o}) \tag{5}$$

$$\mathrm{h_t} = \mathrm{o_t} * \tanh(\mathrm{C_t}) \tag{6}$$

In this way, we decide what parts of the cell state to output by a sigmoid layer as shown in part III.

Experiments

We use python to compose Chinese folk music. We use music21 for MIDI music extraction, use keras to implement LSTM model, and complete the algorithm of folk music composition. The specific process is as follows:

Process MIDI Files. A MIDI file consists of a header chunk and one or more track chunks followed. The head block is marked with MThd, and the track block is marked with MTrk. The header block describes the format of the file, the number of track blocks and so on. The data part of the track block consists of one or more pairs of <delta-time><event>, namely MIDI events.

Learn interactions within music. MIDI files usually contain multiple audio tracks. We will use the following method to learn interactions within music:

(1) If there is only one main melody track, we can perform MIDI files parsing directly.

(2) If the file has multiple tracks, we combine instruments to achieve a fuller sound.

Extract MIDI melody feature. We obtain the key numbers, namely the positioning of music reference sounds: A, B, C, D, E, F, G, beats and the number of audio tracks by the MIDI header file. The pitch and duration are obtained according to the MIDI file encoding rules.

Learn Correlation between Musical Structures. The pentatonic structure has the following characteristics:

(1) There is no Minor Second (also referred to as semitone) between adjacent staff positions.

(2) The distance between adjacent staff positions is the major second or the diminished third.

In addition to providing LSTM with information about metrical structure such as note, chords and velocity, our model also consider intervals by setting constraints for two adjacent pitches. We do not allow semitone between adjacent staff. In this way, our model is able to generate music with more rational structures.

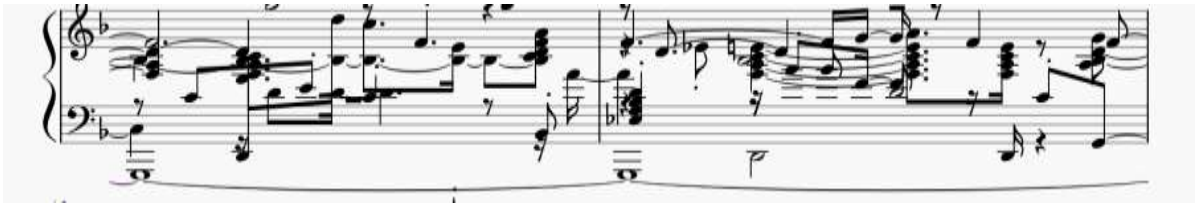


Fig 3 Part of scores without interval constraints

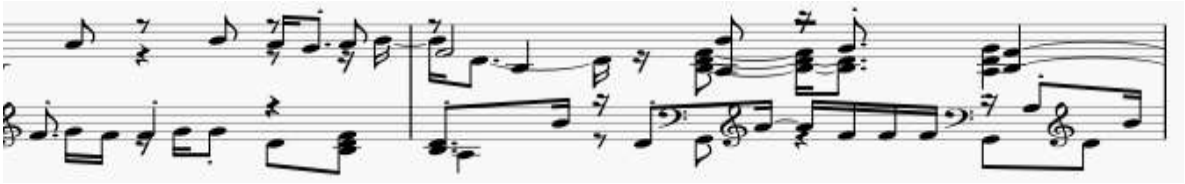


Fig 4 Part of scores applied constraints

Build LSTM Model to Generate Music. Based on the deep learning framework Keras, we set up a 3-layer LSTM model to train music corpus, use dropout to reduce overfitting. The specific steps are as follows:

(1) Obtain related music information from the MIDI file, and extract music corpora information from the obtained grammatical structure.

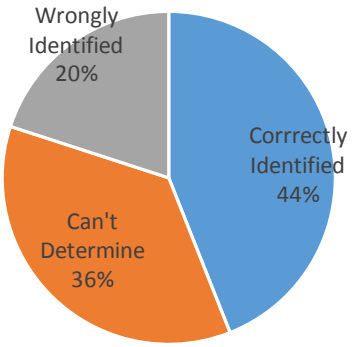
(2) Based on training corpus, establish a 3-layer LSTM. The data is converted to a binary matrix, dropout is set to 0.1, batch-size is set to 128. We use softmax as the activation function and categorical cross-entropy as the corresponding loss function .

(3) Set the number of training rounds to 100 to generate music. And take relevant measures to improve the quality of generated music, such as: smooth processing, delete notes which are too close.

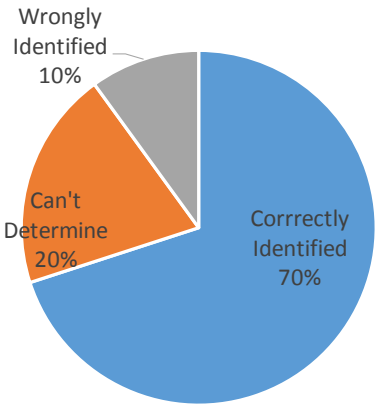
(4) The generated music is stored as midi file and output results.

Evaluation

In order to evaluate the quality of our generated music effectively, we adopt subjective methods. We performed two sets of experiments. Due to the limited conditions, we investigated 50 students and asked them to judge 10 actual human performances and 10 computer-generated songs respectively. The first experiment was performed on short music clips of 10 seconds. Then we set a time limit of 30 seconds to evaluate performance with longer duration. We randomly set the order of the generated and human tracks to reduce bias. The findings are show in Fig. 5(a):



(a)



(b)

Fig 5 Experimental results of short music (a) and performance with longer duration (b)

The results of the first experiment shows about forty percent of students can correctly identified performance between human and computer, while twenty percent tell wrong and about forty percent cannot determine. This concludes that participants could not tell the difference of short music clips between generated and real performances. This demonstrates that our methods for learning Chinese folk music are effective in improving the quality of generated music.

We further analyze the capacity for the model to learn music with longer duration.But the result shows that our model is not sufficient to songs with longer duration. In Fig. 5(b), there are seventy percent of students that can correctly identified human performance. It may result from the model's limited ability to capture more complex information of music architectures.

We also conducted a student survey in which 20 students were asked to listen to 10 pairs of random samples between music with a threshold set to filter mutated notes and music without constraints. Ninety percent of students prefer the first set, saying they are more harmonies. The survey shows that our model can filter out mutated notes and remove noise from music. Participants also made comments that music sounds more colorful with multiple instruments than one instrument alone. This confirms that our model can learn interactions within music.

Due to the limited condition, participants could not notice shifts within a melody from one key to another a semitone higher. Thus we use scores (as shown in Fig. 3 and Fig.4) to show the difference between no interval-limited output and interval-limited output. As shown in Fig. 4, the following scores is more structured than the above one. This proves the usefulness of correlation alignments. The addition of pentatonic structure constraints improved the qualitative output of the model.

Conclusion

Based on the LSTM neural network, our model can learn the correlation of Chinese folk music structure well and generate more harmonious music. Applying the long short-term memory model LSTM to ethnic music creation is beneficial to the development of Chinese music. However, the chords generated by our music are selected from chords which are considered reasonable in training music. They are relatively dependent on the selected music.

Acknowledgements

The research work in this paper was supported by the grants from National Natural Science Foundation of China (61472043) and the Joint Research Fund in Astronomy (U1531242) under cooperative agreement between the NSFC and CAS. Prof. Qian Yin is the author to whom all the correspondence should be addressed.

References

[1] Alpen A. Techniques for algorithmic composition of music. 1995.
[2] C.Ames and M.Domino. Cybernetic composer: An overview. In M.Balaban,K.Ebcioglu, and O.Laske,editors,Understanding Music with AI.AAAI Press, 1992. 186~205.
[3] Ebcioglu K. An expert system for harmonizing chorales in the style of J. S. Bach. In: Balaban M, Ebcioglu K,Laske O, eds.Understanding Music with AI. Cambridge: AAAI Press, 1992. 294−334.
[4] Cope D. Virtual Music: Computer Synthesis of Musical Style. Cambridge: MIT Press, 2001.
[5] Wiggins G, Papadopoulos G, Phon-Amnuaisuk S, Tuson A. Evolutionary methods for musical composition. 1998.
[6] Biles JA. Genjam: A genetic algorithm for generating jazz solos. In: Proc. of the Int'l Computer Music Conf. San Francisco: ICMA,1994. 131−137.
[7] Biles JA. GenJam in transition: From genetic jammer to generative jammer. 2002.
[8] Mozer MC. Neural network composition by prediction: Exploring the benefits of psychophysical constraints and multiscale processing. Cognitive Science, 1994,6:247−280.
[9] Hochreiter S, Schmidhuber J. Long short-term memory. 1996.
[10] Eck D. Finding temporal structure in music: Blues improvisation with LSTM recurrent networks. In: Boulard H, ed. Neural
[11] http://colah.github.io/posts/2015-08-Understanding-LSTMs/
[12] Keunwoo Choi, George Fazekas, and Mark Sandler Text-based LSTM networks for Automatic Music Composition.2016