Research on Chord Generation in Automated Music Composition Using Deep Learning Algorithms

Ming Zhu
Nanchang Institute of Technology, Nanchang, Jiangxi 330000, China
Corresponding address: Ideal Home Community, Honggutan New District, Nanchang City, Jiangxi 330000, China
E-mail: zm4oqw@126.com

Keywords: deep learning, automated music composition, chord generation, transformer

Received: May 24, 2023

With the development of technology, automated music composition has received widespread attention in music creation. This article mainly focuses on the generation of chords in automated music composition. First, relevant music knowledge was briefly introduced, and then the composition of the Transformer model was explained. A two-layer bidirectional Transformer method was designed to generate chords for the main melody and chorus separately, followed by the establishment of chord coloring and sound production models. Ten music professionals and 40 ordinary college students compared the coherence, pleasantness, and innovation of the chords generated by Hidden Markov Model (HMM), Long Short-Term Memory (LSTM), and the method proposed in this paper. The results showed that the chord generated by the method proposed in this paper achieved higher scores in the evaluation. Overall, the scores given by the music professionals and ordinary college students were 3.64 and 3.91, respectively, which were higher than those of the HMM and LSTM methods. The experimental results prove the superiority of the chord generation method proposed in this paper. The method can be applied to automated music composition.

Povzetek: Raziskava se osredotoča na avtomatizirano ustvarjanje glasbe, zlasti generacijo akordov, in predstavi izboljšano metodo z uporabo dvoslojnega dvosmernega pretvornega modela. V primerjavi z metodami HMM in LSTM je nov pristop pokazal višje rezultate v ocenah koherence, prijetnosti in inovativnosti, kar potrjuje njegovo učinkovitost in uporabnost v avtomatizirani glasbeni kompoziciji.

1 Introduction

Music creation requires a high level of knowledge, relevant experience, inspiration, creativity, and other factors for the creator. Therefore, music composition is usually carried out by professional composers with strong expertise, which poses great difficulties for amateur enthusiasts [1]. With the rapid development of technologies such as artificial intelligence, computer-aided music composition, or algorithmic composition, has gradually attracted the attention of researchers [2]. Computer-aided music composition is a method that uses computer technology and combines knowledge from various fields such as mathematics and music to create music and assist musicians [3]. However, this field not only requires the creator to have not only a good foundation in algorithms, but also a certain level of musical knowledge. Therefore, research on automated music composition has become a challenging issue.Music is an essential form of entertainment in people's lives, and artificial intelligence is currently a key area of development [4]. Therefore, automated music composition under the umbrella of artificial intelligence has great development prospects [5]. It not only lowers the threshold of music composition to a certain extent but also provides more musical resources for music lovers. In a piece of music, melody and chord play a very important role, which affects the listenability of the music. At present, there have been many achievements in melody generation in automatic music composition research, while there is little research on chord generation. Therefore, this paper designed a deep learning-based method for chord generation, using a bidirectional Transformer model to generate chords. By analyzing the coherence and pleasantness of the generated chords, the reliability of the method was proved, which is conducive to obtaining more pleasant chord music and makes some contributions to further research on automatic music composition.

2 Related works

The following table shows the current research on automatic music composition.

Literatur e	Method	Result	Limitation
Wang et al. [6]	Deep learning neural network	The composition algorithm based on deep learning can realize music creation, and the qualified rate reaches 95.11%.	The actual music styles are more complex and varied, and how to build a multi-style and multi-thematic music intelligence

		Compared with the composition algorithm in the latest study, the model achieves 62.4 percent satisfaction with subjective samples and a recognition rate of 75.6 percent for musical sentiment classification.	generation model still needs to be explored.
Dua et al. [7]	Recurrent neural network (RNN) with gated recurrent units and long short term memory (LSTM).	The proposed model has a higher signal distortion ratio and increase the accuracy to 78%.	Sheet music generator is still not accurate to the considerable levels and leaves a large room for improvement.
Marsden et al. [8]	Bayesian network and LSTM	The LSTM on average was identified as human-made 36% of the time, while the Bayesian network, on average, had been misidentified 39% of the time.	Limitations in hardware availability lead to limitations in the complexity of the model. The amount of music used for training must be limited, as it may lead to overfitting of the model.
Li et al. [9]	LSTM model	The proposed method is capable of generating new music sequences and smoothly stitching music fragments into one complete audio.	The quality of the compositions depends on the quantity and quality of the audio material.

3 Relevant music knowledge and automatic music composition

The basic attributes of music include:

(1) pitch: the high or low sound of a note, related to the frequency of vibration;

(2) loudness: the volume of a note, related to the amplitude of vibration;

(3) duration: the length of time a note is played, including quarter notes, eighth notes, and sixteenth notes [10];

(4) timbre: the tone color of music, different voices and instruments produce different timbres.

In the music system, C, D, E, F, G, A, and B are pitch names, and do, re, mi, fa, sol, la, and si are roll calls, as shown in Table 1.

Table 1: Numbered musical notation, roll calls, and pitch names

Numbered musical notation	1	2	3	4	5	6	7
Roll call	do	re	mi	fa	sol	la	si
Pitch name	C	D	E	F	G	A	B

Melody is a combination of pitch and rhythm, the soul and foundation of music, and can be divided into vocal and instrumental melodies. Melodies are composed of musical phrases and developed through repetition and variation. Melodies are composed of sections, with each section composed of several phrases, each phrase composed of several measures, and each measure composed of several motifs. Motifs are the smallest unit of a melody.

Chords are a combination of two or more (usually three) notes and can be divided into triads, seventh chords, and so on. depending on the number of notes. Triads are the most common type of chord, as shown in Figure 1.

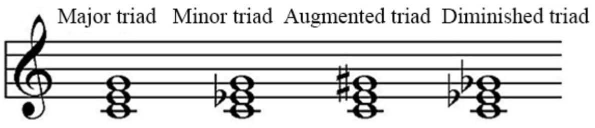

Figure 1: Triads

Harmony is the combination of two or more voices, one of which is chords, which refers to the vertical movement of harmony, and the other is harmony progressions, which refers to the vertical movement of harmony created by connecting chords.

In automatic music composition, the following methods are currently used.

(1) Markov chain: Notes are selected through an n-dimensional transition table to generate melodies, but training on a large dataset is required. It cannot learn abstract concepts in music.

(2) Genetic algorithm: Information is stored by taking melodies or motifs as chromosomes. Regeneration and selection are performed to create new melodies, but the efficiency in composition is low due to the strong influence of subjectivity.

(3) Music rules: A knowledge base containing various music rules is used to create music content, but it is influenced by existing human thinking.

(4) Deep learning: e.g. RNN, LSTM, etc., and training the neural network can learn music rules and generate relevant sequences.

4 Chord generation method based on Transformer model

4.1 Transformer model

Currently, Hidden Markov Models (HMMs) are commonly used in chord generation, where the melody is used as an observation value to predict the corresponding chord. In music, there is a long-term dependency between chords and melody, but in HMMs, the current state is only influenced by previous states. The Transformer model, as a sequence generation model [11], has better learning performance compared to RNNs, LSTMs, etc., due to the addition of self-attention mechanism, and has shown good performance in natural language processing, machine translation, etc. [12]. Therefore, this paper studies chord generation using the Transformer model. First, a brief introduction of the Transformer model is presented.

(1) Self-attention mechanism

Self-attention is the most critical component of the Transformer model. It is assumed that there are query vector q, key vector k, and value vector v, which are processed into matrices Q, K, V. The calculation process of the self-attention mechanism is:

$$Attention(Q,K,V) = softmax\left(\frac{Q^TK}{\sqrt{d_k}}\right)V, \quad (1)$$

where d_k is the dimension of vector k.

(2) Multi-head attention

Composed of multiple self-attentions, it can effectively improve the training speed of the model. The calculation formula is:

$$MultiHead(Q,K,V) = Concat(head_1, head_2, \cdots, head_h)W_O, \quad (2)$$

$$head_i = Attention\left(QW_i^Q, KW_i^k, VW_i^V\right). \quad (3)$$

(3) Position-embedding

It is used for capturing the location information of a sequence, and its calculation formulas are:

$$PE(p,2i) = \sin\left(\frac{p}{10000^{2i/d_{model}}}\right), \quad (4)$$

$$PE(p,2i+1) = \cos\left(\frac{p}{10000^{2i/d_{model}}}\right), \quad (5)$$

where p is the position index and d_{model} is the dimension of a vector.

The Transformer model has an encoder-decoder structure. The encoder includes a multi-head attention and a Feedforward Neural Network (FNN), while the decoder includes a multi-head attention, an attention mechanism from the encoder to the decoder, and an FNN. In this structure, if the input is X^t, it is transformed into a hidden state H^t after passing through the encoder. After decoding, the output is Y^t. The calculation formulas of multi-head attention and FNN are as follows:

$$MultiHead'(Q,K,V) = LayerNorm(MultiHead(Q,K,V) + Q), \quad (6)$$

$$FFN'(X) = LayerNorm(FFN(X) + X), \quad (7)$$

where X is a sandwich matrix.

4.2 Chord generation method

Chords have similar characteristics to melodies, are influenced by the chords before and after, and have characteristics such as sequential and repetitive. For the chord direction generation, in order to fully learn the information of before and after chords, this paper designs a bidirectional Transformer model to learn the information before and after the current state respectively. Then, since music generally adopts the structure of verse and chorus, the verse and chorus have large differences in melody. Therefore, this paper uses two bidirectional Transformer models to generate the chords of the verse and the chorus, respectively. In addition, there is also an articulation between the verse and the chorus, so a self-attention mechanism is added to the verse chord generation model to learn the articulation between the verse and the chorus.

The structure generation for chords is divided into two main elements.

First, for chord coloring, i.e., the pitch composition of a chord, it is assumed that its input includes chord sequence $\{x_i\}_{i=1}^T$ and duration sequence $\{d_i\}_{i=1}^T$. Root sequence $\{b_i^c\}_{i=1}^T$ and pitch sequence $\{p_i^c\}_{i=1}^T$ need to be predicted, where c represents coloring. Then, the input embedded input is written as: $e_i^c = W^{e^c}(d_i x_i)$, the sequential modeling layer is $h_i^c = f^c(e_i^c | e_{1:T}^c)$, and the predicted results are $p_i^c = sigmoid\left(W^{p^c} h_i^c\right)$, $b_i^c = softmax\left(W^{b^c} h_i^c\right)$, where W^{e^c}, W^{p^c}, and W^{b^c} are learnable parameters.

The sequential modeling layer is composed of multi-head attention, and its calculation mode is as follows:

$$h_i^{c(l)} = W^{outer}\sigma(W^{inner}u_i + b_1) + b_2,$$

$$u_i = W^u\left(u'_{i1} \oplus u'_{i2} \oplus \cdots \oplus u'_{iJ}\right) + h_i^{c(l-1)}, \quad (8)$$

$$u'_{ij} = V_j softmax\left(\frac{K_j^T q_{ij}}{\sqrt{d}}\right), \quad (9)$$

$$q_{ij} = W_j^Q h_i^{c(l-1)}, \quad (10)$$

where l stands for the iteration step, 2 here, σ stands for RELU activation function, W^{outer}, W^{inner}, and W_j^Q are learnable parameters, and J is the number of heads in multi-head attention, 8 here. The loss function is:

$$\tau^c = \sum_{i=1}^T\left[BCE\left(p_i^{c\prime}, p_i^c\right) + CCE\left(b_i^{c\prime}, b_i^c\right)\right], \quad (11)$$

where BCE represents a binary cross entropy, CCE represents a classification cross entropy, and $p_i^{c\prime}$ and $b_i^{c\prime}$ are real numbers of p_i^c and b_i^c.

Then, for chord voicing, i.e., spacing between the pitches of a chord and repetition, it is assumed that its inputs are root sequence $\{b_i^\tau\}_{i=1}^T$, pitch sequence $\{p_i^\tau\}_{i=1}^T$, and duration sequence $\{d_i\}_{i=1}^T$, and v is voicing. Chord voicing $\{v_i\}_{i=1}^T$ needs to be predicted. The model also uses the same structure as the chord coloring:

$$e_i^v = W^{e^v}\left(d_i(p_i^v \oplus b_i^v)\right),\ h_i^v = f^v(e_i^v | e_{1:T}^v),\ v_i = sigmoid(W^v h_i^v),$$

where W^{e^v} and W^v are learnable parameters. If the target voicing of the i-th chord is v'_i, then its loss function is:

$$\tau^v = \sum_{i=1}^T BCE(v'_i, v_i). \quad (12)$$

The output of the coloring model is used as input to the vocalization model to generate a sequence with a chord structure based on the coloring post sequence.

5 Experiment and analysis

One hundred pieces of classical guitar music were collected for the experiment. To verify the effectiveness of the chord generation method designed in this paper, only the melody and guitar chords were retained. At the same time, the cross-repetition of the verse and chorus of the songs was removed, and only one section of the verse and one section of the chorus were preserved for the experiment. The piano roll representation method was used to represent the chords, where the value was 1 if the pitch was included in the chord structure, and 0 otherwise. For example, for the G chord in the key of C major (Figure 2), the chord representation is shown in Table 2.

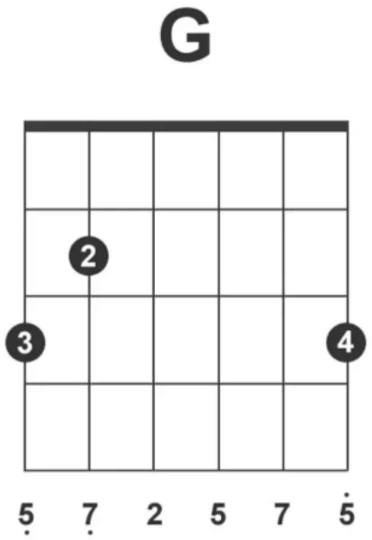

Figure 2: The G chord in the key of C major.

Table 2: Examples of chord representation

……	F	#F	G	#G	A	#A	B	C	#C	D	#D	E	F	……
……	0	0	1	0	0	0	1	0	0	1	0	0	0	……

During preprocessing, the chords were split into eights bars. The repetition number of the bidirectional Transformer model was set to 0, the dimension was set as 128, the initial learning rate was set as 0.0001, and the batch size was set to 16. There is currently no scientific and objective evaluation method for automatic music composition, so subjective evaluation methods are usually used. Fifty people participated in the evaluation, including ten music professionals who was major in music and have experience in music composition, and 40 ordinary college students. The scoring system was a five-point scale, with 5 being the highest and 1 being the lowest. The scoring criteria are as follows:

(1) chord coherence: the degree of coherence of chord transitions;

(2) chord pleasantness: the listenability of the chords;

(3) chord creativity: the creativity and innovation of the generated chords.

To demonstrate the superiority of the chord generation method proposed in this paper, it was compared with the HMM method [13] and the LSTM method [14], and the results are shown in Table 3.

Table 3: Chord scoring results

		HMM	LSTM	The Transformer-based method
Music professionals	Chord coherence	2.1	2.5	3.6
	Chord pleasantness	2.2	2.7	3.8
	Chord creativity	1.8	2.3	3.5
Ordinary college students	Chord coherence	3.1	3.5	3.9
	Chord pleasantness	3.2	3.7	4.0
	Chord creativity	3.3	3.6	3.8

From Table 3, first of all, in terms of the scores given by music professionals and ordinary college students, the scores given by ordinary college students were slightly higher than those given by music professionals. This may be because the professionals have a deeper understanding of musical knowledge, so they tend to give harsher evaluations from a professional perspective when evaluating chords, resulting in lower scores. Secondly, when comparing the chord scores generated by the HMM, LSTM, and Transformer-based methods, the chord scores generated by the Transformer method were significantly higher than those generated by the HMM and LSTM, and music professionals and ordinary college students have made consistent evaluations. Music professionals felt that the chord generated by the HMM had poor creativity, so they gave it an average score of only 1.8. The scores for coherence and pleasantness were also not high. The chords generated by the LSTM performed slightly better than those generated by the HMM, but their coherence, pleasantness, and creativity scores were not higher than 3.0 points. The Transformer-based method scored 3.5 for creativity, and 3.6 and 3.8 for coherence and pleasantness, respectively, which were significantly higher than the HMM and LSTM methods. Although the scores for the the HMM- and LSTM-generated chords given by ordinary college students were slightly higher than the scores given by music professionals, there was a gap compared with the scores of the chord generated by the Transformer-based method. A score of 4.0 was given for the pleasantness of the chord generated by the Transformer-based method, proving that the chord generated by the proposed method had good listenability.

Further analysis of the evaluation results was conducted using the weighted average method. The scores for coherence, pleasantness, and creativity were combined in a ratio of 5:3:2. The scores given by music professionals and ordinary college students were combined in a ratio of 6:4. The final evaluation results are shown in Figure 3.

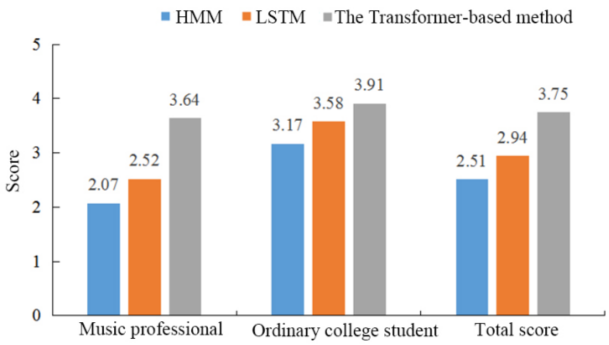

Figure 3: The comprehensive comparison of three chord generation methods.

From Figure 3, first of all, after combining the three scores, music professionals gave a score of 2.07 for the chord generated by the HMM, a score of 2.52 for the chord generated by the LSTM, and a score of 3.64 for the chords generated by the Transformer-based method. The score of the Transformer-based method was 1.57 points higher than that of the HMM method and 1.12 points higher than that of the LSTM. Ordinary college students gave a score of 3.17 for the chord generated by the HMM, a score of 3.58 for the chord generated by the LSTM, and a score of 3.91 for the chord generated by the Transformer-based method. The score of the Transformer-based method was 0.74 points higher than that of the HMM and 0.33 points higher than that of the LSTM. This indicated that the chord generated by the proposed method was of higher quality from the perspective of both music professionals and ordinary listeners. Finally, in terms of the total score, the total score for the chord generated by the HMM was 2.51, and the total score for the chord generated by the LSTM was 2.94. The total score for the chords generated by the Transformer-based method was 3.75 points, which was 1.24 points higher than the HMM and 0.81 points higher than the LSTM. These results demonstrated the reliability of the Transformer-based method.

6 Discussion

The development of computers has driven the innovation of music technology, and automatic music composition has been rapidly developed and applied to various fields of music production. However, the current automatic music composition still faces some limitations. The creation of complex music is still difficult due to the lack of emotion and creativity of human artists, and moreover the algorithmic models all rely on a large amount of training data to achieve high-quality compositions. Deep learning methods are able to adapt to different data mining tasks through self-learning and parameter adjustment, and have significant advantages in processing high-dimensional and complex data, so they also have good applications in the field of automatic music composition. In this paper, based on deep learning, a two-layer bidirectional transformer chord generation model was designed and experimented with guitar music as an example.

According to the results, the chords generated by the bidirectional transformer after learning have achieved good results in the subjective evaluation. Compared with HMM and LSTM, the two-layer bidirectional transformer was more adequate and complete in learning musical features, and therefore the generated chords were closer to the results of human compositions and achieved higher scores after being evaluated by music professionals and general college students. Specifically, the chord innovation score was lower than the chord coherence and chord pleasing scores, which indicates that, similar to the results of the current study, the chords obtained by the algorithm are still deficient in terms of innovative compositions.

The research in this paper mainly focuses on chord generation. This paper separated the melody and chords of music, breaking through the current shortcomings of automatic music composition in chord generation, and obtained more reliable results, making some contributions to the progress of automatic music composition. In future research, the study of human music theory and artistic creation will be strengthened to further improve the algorithm's ability to learn musical tonality and emotion, so that the algorithm can simulate the thinking and creation process of human music and improve the innovation of automatic music composition, thus promoting the rich and diverse development of automatic music composition.

7 Conclusion

In this paper, a two-layer bidirectional chord generation method was designed using the Transformer model in deep learning to generate chords for both the verse and chorus sections of a song. Experimental analysis revealed that the chords generated by the Transformer-based method exhibited better performance in terms of coherence, pleasantness, and creativity compared to the HMM and LSTM methods. Both music professionals and ordinary college students gave higher scores for the chord generated by the proposed method, demonstrating the effectiveness of the method. This method can be further promoted and applied to practical automatic music composition.

References

[1] Itou K, Tanaka D (2018) Automatic Electronic Organ Reduction System Based on Melody Clustering Considering Melodic and Instrumental Characteristics. *2018 IEEE International Symposium on Multimedia (ISM)*, pp. 151-158.

[2] Brook T (2020) Musicking with Music-Generation Software in Virtutes Occultae. *Leonardo Music Journal*, 30, pp. 3-7. https://doi.org/10.1162/lmj_a_01086

[3] Mukherjee H, Dhar A, Ghosh M, Obaidullah S M, Santosh K C, Phadikar S, Roy K (2020) Music chord inversion shape identification with LSTM-RNN. *Procedia Computer Science*, 167, pp. 607-615. https://doi.org/10.1016/j.procs.2020.03.327

[4] Navarro M, Corchado J M (2018) Machine Learning in Music Generation. *Oriental Journal of Computer Science and Technology*, 11, pp. 75-77.

[5] Kaliakatsos-Papakostas M, Floros A, Vrahatis M N (2020) Artificial intelligence methods for music generation: a review and future perspectives - ScienceDirect. *Nature-Inspired Computation and Swarm Intelligence,* 2020, pp. 217-245. https://doi.org/10.1016/B978-0-12-819714-1.00024-5

[6] Krishnan V P, Rajarajeswari S, Krishnamohan V, Sheel V C, Deepak R (2020) Music Generation Using Deep Learning Techniques. *Journal of Computational and Theoretical Nanoscience*, 17, pp. 3983-3987. https://doi.org/10.1166/jctn.2020.9003

[7] Dua M, Yadav R, Mamgai D, Brodiya S (2020) An Improved RNN-LSTM based Novel Approach for Sheet Music Generation. *Procedia Computer Science*, 171, pp. 465-474. https://doi.org/10.1166/jctn.2020.9003

[8] Gioti A M (2020) From Artificial to Extended Intelligence in Music Composition. *Organised Sound*, 25, pp. 25-32. https://doi.org/10.1017/S1355771819000438

[9] Li X F, Feng T T, Luo S, Zhang X L (2018) Automatic music composition algorithm based on recurrent neural network. *Jilin Daxue Xuebao (Gongxueban)/Journal of Jilin University (Engineering and Technology Edition)*, 48, pp. 866-873. https://doi.org/10.13229/j.cnki.jdxbgxb20170509

[10] Temperley D (2021) The origins of syncopation in American popular music. *Popular Music*, 40, pp. 18-41. https://doi.org/10.1017/S0261143021000283

[11] Popel M, Bojar O (2018) Training Tips for the Transformer Model. *Prague Bulletin of Mathematical Linguistics*, 110, pp. 43-70. https://doi.org/10.2478/pralin-2018-0002

[12] Maclean A, Wong A (2021) Where do Clinical Language Models Break Down? A Critical Behavioural Exploration of the ClinicalBERT Deep Transformer Model. *Journal of Computational Vision and Imaging Systems*, 6, pp. 1-4. https://doi.org/10.15353/jcvis.v6i1.3548

[13] Ikesu R, Taguchi A, Hara K, Kawana K, Tsuruga T, Tomio J, Osuga Y (2022) Prognosis of high-risk human papillomavirus-related cervical lesions: A hidden Markov model analysis of a single-center cohort in Japan. *Cancer Medicine*, 11, pp. 664-675. https://doi.org/10.1002/cam4.4470

[14] Hui L I, Wang R, Wang C (2022) Prediction of Partial Ring Current Index Using LSTM Neural Network. *Chinese Journal of Space Science*, 42, pp. 873-883. https://doi.org/10.11728/cjss2022.05.210513061