A Review of Research on Emotional Music Generation

Yuhang Wang*

College of Science, School of Jiangsu University of Science and Technology, Zhenjiang, China

*Corresponding author: 222210506207@just.edu.cn

Abstract. With the expansion of the development field of artificial intelligence, emotional music generation's potential in the field of music generation has been gradually explored. In this paper, rule-based, deep learning and reinforcement learning methods for emotional music generation are analyzed, respectively. Rule-based method generates music according to certain theories and patterns, which is interpretable but has a single style. Deep learning generates music by learning the characteristics of music data through neural network models, which have rich emotional levels but rely on a large amount of data and have poor flexibility. Reinforcement learning allows agents to optimize behavior strategies in interaction with the environment to generate music, which is flexible but complex. Through comparative analysis, it is found that the rule-based method can be used as a basic framework to constrain other methods, deep learning has outstanding performance in creating diversity and generation quality, and reinforcement learning has obvious advantages in generating specific emotions. The integration of the three methods is expected to further improve the effect of emotional music generation and provide strong support for music creation.

Keywords: Emotional music generation; Rule-based; Deep learning; Reinforced learning.

1. Introduction

As an important carrier of human emotions, music is filled with all aspects of our lives, such as school radio stations, BGM accompanied by TV dramas, short videos, games and other pictures. The right music is key to enhancing the experience. With the rapid development of society, the improvement of material level makes people's pursuit of spiritual satisfaction in art and culture show a significant growth trend. Emotion is a manifestation of the spiritual level. People produce and experience emotions all the time. Music also contains emotions and can express emotions. The same classification of emotions applies to music. The commonality and close connection of the two are extremely similar. Based on the fact that music is one of the direct and critical paths for human emotional resonance and spiritual dependence, how to better capture and analyze personal emotions and generate real and high-quality music samples close to actual experience to meet personalized needs is a topic worthy of further development. In recent years, artificial intelligence has been brilliant in the practical application of various fields. It not only provides convenience to greatly improve work efficiency, but also reflects the autonomy and creativity. Therefore, people do not stop at the scope of engineering, and actively broaden to the scope of art. Through a certain algorithm combined with model building let machine learning. Nowadays, the continuous and stable development of deep learning technology empowers the generation of emotional music. ejaren Hiller and Leonard Isaaeson achieved the first music composed entirely by computer in 1957, which started the research of computer music generation. In 2018, Huang et al. first applied Transformer to the field of symbolic music generation [1]. Wang Zhongmin et al. calculated the Phase Lock Value (PLV) between the features of multi-channel EEG signals to construct graphical signals, and the implicit information between different frequency bands could be utilized [2]. Song Tengfei et al. proposed a graph convolutional model DGCNN. During the training process of the model, the parameter values of the matrix are dynamically optimized and updated to explore the connection relationship between EEG electrodes and obtain more discriminative emotional features to achieve high emotion recognition accuracy [3]. In 2020, Zhong et al. proposed a regularized graph neural network RGNN that considers local connectivity and global channel relationship to design the adjacency matrix. When constructing the adjacency matrix, the asymmetric information between the left and right brain is taken into account, and 9 pairs of symmetric connectivity relationships are initialized in the adjacency matrix

[4]. Du et al. proposed a multi-dimensional graph convolutional network, in which the spherical distance relationship between electrode points was used to construct the adjacency matrix, and some important points were manually selected to construct the adjacency matrix. The two matrices were normalized to [0,1] and added to the average [5]. In 2021,Hung et al. constructed EMOPIA, a dataset of emotional piano music [6]. In the same year,Alabdulkarim et al. used PPO reinforcement learning algorithm to optimize GPT-2 generative pre-trained language model [7]. Just being able to generate music is not enough. At present, researchers have been in the stage of trying to optimize the music generation, and gradually establish the human-led, directional and targeted generation of high quality emotional music that meets the needs. This paper will analyze and review the general architecture of emotional music generation based on rules, based on deep learning and based on reinforcement learning, as well as the characteristics, advantages and disadvantages of representative model structures and their different application fields. Finally, through the analysis and summary of the current situation, we will discuss different ideas and put forward prospects for the future.

2. Overview of Related Technologies

2.1. Rule-based Emotional Music Generation

In the early stage, the EEG signals of each band generated by emotion induction are collected and preprocessed. Differential entropy is used for feature extraction after obtaining valuable EEG signals. The emotion recognition model design is divided into seven layers, the first layer is the input layer, that is, the features extracted from the input EEG signals. The second layer uses the distance of the spatial nodes of EEG to construct the adjacency matrix representing the topology of the EEG channels, which is the key for sentiment classification. The third and fourth layers are graph convolution layers, which extract EEG signal features from different scales respectively. The fifth layer is the fusion layer, which fuses the previously obtained features to obtain new features. The sixth layer is the channel attention layer, which uses the adaptive channel mechanism to assign weights to different channels according to their importance to obtain a new feature matrix. The seventh layer is a fully connected layer, and the features obtained in the previous layer are flattened to one dimension and then connected to the layer and the softmax function classifier. The music generation method is based on the Transformer model combined with parameter mapping [8]. Each chord type has its associated emotional label, the encoder takes a series of high-level musical parameters as conditional input, and the decoder outputs a generated sequence of leading tables for each measure. After obtaining the emotional information, the valence and arousal parameters are mapped to five realized music structure parameters, namely rhythm, rhythm complexity, loudness, pitch and chord, according to the mapping rules, and the brain music is generated by these parameters. Gradually varying the input parameters valence and arousal is able to generate a continuous stream of music with a specific EE quality.

2.2. Emotional Music Generation with Deep Learning Transformer Models

The former model is composed of a multi-layer encoder and a decoder, and it is a time series model that relies on the attention mechanism. Discrete-emo Transformer model is an emotional music generation model based on Discrete emotional datasets [1]. The model uses the fragment recursion mechanism in the Transformer-XL model and the relative position encoding to process the long temporal sequence of the model. The fragment recursion mechanism divides the long sequence into equal-length fragments, and expands the stored hidden sequence as information to realize fragment interaction when processing the next fragment, which improves the accuracy and logicality of the generated content. Relative position encoding focuses on relative position rather than absolute position to enhance the adaptability of the model. This model uses Flash attention mechanism instead of self-attention mechanism. The block processing was adopted, the characteristics of different memory in GPU were used for operation, and the stored data were re-calculated to optimize the memory and improve the calculation efficiency. The Continuous-Emo Transformer model is an emotional music generation model based on Continuous emotional datasets [9]. The backbone network of the model is mainly composed of multi-layer self-attention layer, feed-forward layer, regularization layer, and linear layer. The fragment recurrence mechanism and relative encoding adopted by the model are the same as that of the Discrete-Emo Transformer model. In addition, the model uses the emotion enhancement module to fuse the multi-feature of the music element, convert the low-level emotional label to the high-level emotional label, introduce the attention internal conditional control mechanism, strengthen the ability of the model to learn emotional features, and guide the model to generate emotional music in line with expectations. Transformer model is shown in Fig.1.

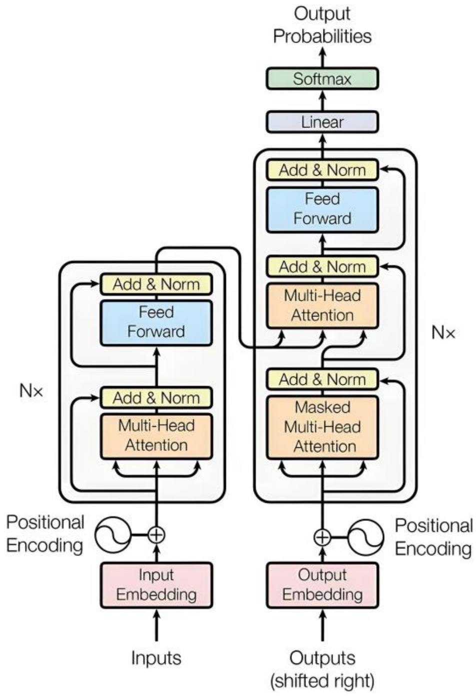

Fig. 1 Transformer model diagram [9]

2.3. Emotional Music Generation with Reinforcement Learning

Reinforcement learning guided emotional music generation model is mainly divided into two parts: generation response and evaluation optimization. The generation response phase uses GPT-2 as the backbone model [10]. The GPT-2 model architecture is mainly based on the decoder part of the Transformer, which learns rich features and information through a large number of parameters and multi-layer structure to achieve strong generation ability. In the evaluation optimization stage, the reward value obtained by the music emotion classification model is input into the PPO reinforcement learning algorithm to guide the model to optimize the generation strategy, and the iterative optimization is continued until convergence, and then the emotional music in line with expectations is generated.

3. Analysis Of Experimental Results

3.1. EMOPIA Dataset

The EMOPIA Dataset is a multimodal popular piano dataset for emotion recognition and emotive-based music generation, containing 1087 music clips taken from 387 songs with both audio and MIDI modal data. The paper have fragment-level emotion labels annotated by four professional annotators. Based on Russell's valence-arousal emotion space, the emotion is divided into four quadrants: high arousal high positive valence (happiness), high arousal low positive valence (anger), low arousal low positive valence (sadness), and low arousal high positive valence (relaxation). It has the characteristics of multi-modal characteristics, rich emotional annotation, and moderate data scale.

It is mainly used in music emotion recognition, music classification, music information retrieval, music generation, music style transfer and other research fields related to music emotion.

3.2. Analysis

Based on the rules of emotional music, the rules of the music generation are about 75 percent, and the rule coverage can reach 60 percent, and the user satisfaction questionnaire shows that the average 65 percent of the users' recognition is expected to be expected.

The classification accuracy of emotional music generation style based on deep learning can reach 70%-85%, the feature distance between different music samples is about 0.4-0.6, and the average coherence score is 6-8.

The cumulative reward of emotional music generation based on reinforcement learning was stable in the range of 80-120 after multiple iterations, and the exploration was first high and then low, which gradually decreased from 80% to about 20%. The convergence speed of the strategy reached an average of 500-1000 iterations.

The process of generating music is simple and efficient, which meets the requirements of rapid generation, but the generated music emotion type is single, the structure and the pattern are fixed, the more mechanical, the lack of the level, the change needs to be artificial adjustment, the creativity is not strong adaptability and the flexibility is poor. Different emotional types, creative and creative, have a high degree of creativity and emotional expression, and the model structure can be adjusted independently, and there is a certain adaptability, and the speed of production is guaranteed, but because of the complexity of the structure, the calculation of the resources is large, the time is long, and the flexibility is limited by the large-scale training data requirements. In the long-term interaction, the emotional music of music, which generates the desired emotion, is high, and can adjust the generation strategy according to the environmental feedback dynamic adjustment, and the flexibility is strong, but the training requires a lot of interaction and trial error, the process is complicated, the process is complicated, the calculation cost is high, the training time is long, the production is low.

4. Conclusion

Emotional music generation aims to produce music that conforms to the expected emotion through technical models. By analyzing the framework of emotional music generation models of three different technologies, this paper summarizes their advantages and disadvantages, that is, rule-based generation has strong interpretability but a single style, deep learning has diverse styles but relies on data and has high difficulty in data acquisition, reinforcement learning meets the needs but has high difficulty in training. Music transmits emotions through the combination and collision of melody, rhythm and other musical elements, and different music styles have their own unique emotional colors. Current technology is difficult to accurately capture and express delicate emotions, and the coordination of the integration of multiple styles needs to be considered. High-quality emotional music data is expensive to obtain and insufficient in quantity, which is more likely to cause copyright disputes. The degree of acceptance of emotion varies from person to person, and different emotional experiences bring difficulties to the improvement of the model. Based on this, the technical breakthrough at this stage can focus on the individual optimization based on different rules and algorithms, such as the optimization of Transformer architecture, the optimization of reinforcement learning algorithm, and the optimization and fusion application of mapping rules. Combining text, image, video and so on, adopting multi-modal fusion to assist emotional music generation can also improve the emotional expression of music. Emotional music generation is currently mainly used in film and television, games, and music therapy, which has achieved remarkable results but is also beset with difficulties, such as the accuracy of expression needs to be improved, data acquisition quality problems and copyright disputes. Multimodal fusion is the general trend, the application field is gradually expanding, and the value of emotional music generation is expected.

References

[1] Huang C Z A, Vaswani A, Uszkoreit J, et al. Music transformer: Generating music with long-term structure. In Proc. Int. Conf. Learn. Represent. 1-14, 2014.

[2] Wang M, El-Fiqi H, Hu J, et al. Convolutional neural networks using dynamic functional connectivity for EEG-based person identification in diverse human states. IEEE Transactions on Information Forensics and Security, 2019, 14(12): 3259-3272.

[3] Song T, Zheng W, Song P, et al. EEG emotion recognition using dynamical graph convolutional neural networks. IEEE Transactions on Affective Computing, 2018, 11(3): 532-541.

[4] Zhong P, Wang D, Miao C. EEG-based emotion recognition using regularized graph neural networks. IEEE Transactions on Affective Computing, 2020, 13(3): 1290-1301.

[5] Du G, Su J, Zhang L, et al. A Multi-Dimensional Graph Convolution Network for EEG Emotion Recognition. IEEE Transactions on Instrumentation and Measurement, 2022, 71: 1-11.

[6] [6] Hung H T, Ching J, Doh S, etal. EMOPIA:A multi-modalpop piano dataset for emotion recognition and emotion-based musicgeneration // Proceedingsofthe 22nd International Society for Music Information Retrieval Conference. Online: ISMIR,2021:318-325.

[7] Alabdulkarim A, Li W, Martin L J, et al. Goal-directed story generation: Augmenting generative language models with reinforcement learning [EB/OL].(2021-12-16)[2023-01-27]. https: // arxiv.org/pdf/ 2112.08593

[8] Heng W. Based on the figure of the neural network electrical brain music emotion recognition and generate (master's degree thesis, Beijing jiaotong university). https://kns.cnki.net /kcms2/article/abstract?v=NK8hpUzgeRUg4JMN1q6WyIjpU 2023, euLh4fHS-JDybN5IHT6 mp1hxhjeelNq50gxraMeTKJHcfTlpPDHtD8XpTuF6lHJaqF8gd5BxkricHRSqIrkmlgPBz0fHMNx4FJ7 PgdK&uniplatform=NZKP T&language=CHS

[9] Wang X H. Research on Transformer-based emotional music generation (Master thesis, 2005). Wuhan University of Light Industry). 2024.

[10] Shen Z X, Xie X L, Yin H, Yang L & Li H F. Affective Music Generation with Reinforcement Learning Guided Pre-trained Models. Journal of Fudan University (Natural Science Edition)2024, (03),336-343. doi:10.15943/j.cnki.fdxb-jns.20240605.001.