Music Composition Based on Generative Modeling with Artificial Intelligence: Challenges, Innovations and Perspectives

Zichen Wang *

Brunel London School, North China University of Technology, Beijing, China

* Corresponding Author Email: 21190010108@mail.ncut.edu.cn

Abstract. In today's era of rapid technological advancement, the emergence of artificial intelligence has brought new opportunities to the field of music creation. Music, as an important part of human culture, has expressed emotions in its own unique way since ancient times and has evolved and innovated over time. This paper introduces the progress made by artificial intelligence in music creation, describes the application and innovation of multi-granular attention Transformer, Variable Auto-Encoder (VAE), and Generative Adversarial Network (GAN) and other technological models in the field of music generation and creation, objectively analyzes the current technological bottlenecks and the dilemmas of artistic integration, and looks forward to the future development trend of the technology and the prospects of its application. The future development trend of the technology and the prospect of its application will be analyzed objectively. Through comprehensive and in-depth research, the paper will sort out the development vein of the field, provide references for subsequent related research, and promote the further integration and development of the field of music creation and artificial intelligence to meet people's growing cultural and artistic needs.

Keywords: Artificial Intelligence, Music Generation, Generative Model, Challenges and Prospects.

1. Introduction

Music plays a large part in human life and is an important piece of the puzzle that makes up human culture. Throughout history, music has had an unbroken legacy of iterative and innovative styles that express feelings or tell a story in a non-literal form. As technology advances, the way in which music is created is also changing, from the initial strumming of keys and pen-and-paper scores to today is computerized arrangements of scores, and the emergence of artificial intelligence has undoubtedly opened up even more possibilities for music creation.

The emergence of artificial intelligence generative models to provide inspiration for music creation, it can learn different styles of music data, mimic the pattern of the music, and quickly generate music fragments, greatly reducing the time spent in finding inspiration and writing notes, the model can generate its own initial music draft, on which the musician optimizes. Modeling can also explore new musical styles and inspire more innovative compositions. At the same time, the market demand for soundtracks for different scenarios is increasing, and the model can provide more accurate music by quickly matching the right music to the provided scenarios and emotions.

In terms of modeling innovation, Zhang et al. proposed the idea of generating high-quality symbolic music with a fine-grained discriminator to generate higher quality music [1]. Focusing on musical structure, Souza et al. discuss the effects of different note duration representations on the musical Transformer, comparing the advantages and disadvantages of the representations in building musical structure [2]. Wang et al. is MeloTrans model generates music based on human creative habits, and studies the patterns and laws of human creative music, so that the generated music is more in line with human creative logic [3]. Kundu et al. carried out a study on emotion-guided image-to-music generation, which utilizes emotional information in images to drive music generation, realizing the association between music and emotion to satisfy a specific emotional atmosphere [4]. Epure et al. attempted to utilize high-level song descriptors to implement natural language-based music recommendation, mining high-level semantic features of songs to recommend music that better matches needs and preferences [5].

The purpose of this paper is to explore the developments, challenges, achievements, and perspectives for the future in this field. Analyze the research results and sort out the development of the field of AI music composition, including the innovations of various generative models such as Transformer, Auto-Encoder (VAE), and Generative Adversarial Network (GAN) in the areas of melody, harmony, multi-tracking, and stylistic adjustments in music composition. In-depth analysis of the challenges that these innovations may face in the process of development, such as the further improvement of music quality, compatibility of the integration of different technologies and other issues, and on the basis of these prospects for the future direction of the development of AI music composition.

2. Different specific techniques/models

2.1. Music Generation Techniques Based on Multi-Granularity Attention Transformer

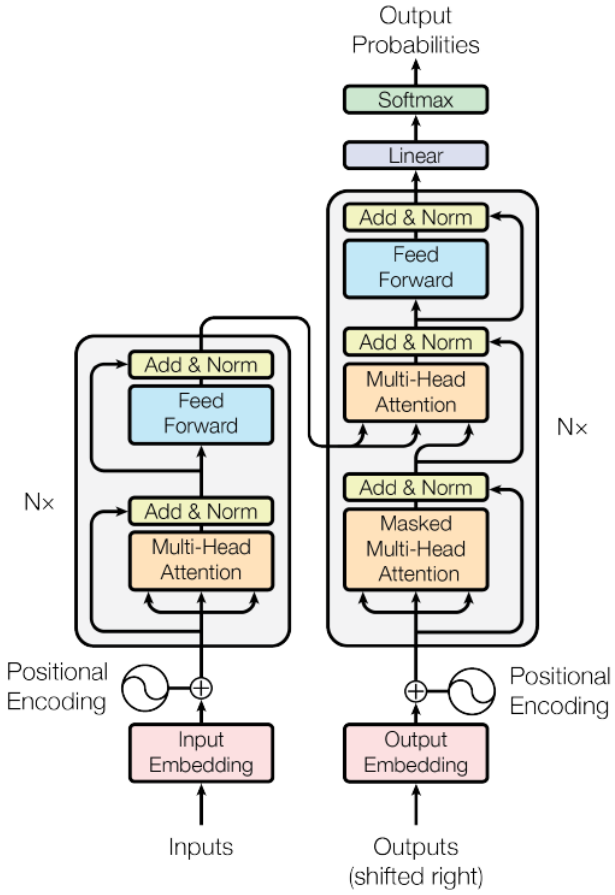

Figure 1. Transformer structure diagram[6]

In the context of music generation, the Transformer model architecture in Fig. 1 is able to capture complex structures in music, such as melodic fluency and rhythmic regularity, through its elaborate multi-head attention mechanism.

The introduction of a multi-granularity attention mechanism to improve the traditional Transformer model makes it uniquely innovative in the field of music generation.

While traditional Transformer models focus equally on the entire sequence with the same importance, thus making it difficult for the model to learn such structural features as music, the core of the multi-granularity attention mechanism is to capture different levels of features in music data. Utilizing fine-grained attention to focus on bar structure, e.g., the study found that a bar in the music was similar to the previous bar 2, bar 4, or multiples bars, and captured the interrelationships of the musical structure by focusing on these bars. Use coarse-grained attention to focus on other bars for additional contextual information. The specific operation first embeds the sequence into the vector

space through the embedding layer and performs linear projection, then the multi-granularity attention layer models the music contextual relationship, and finally predicts the next marking with Softmax. When generating music, the model can rely on this mechanism to grasp the melodic direction, the rhythmic change pattern and so on.

When processing musical elements, the model learns repetitive melodic fragments and variation patterns so that the resulting melodies are coherent and varied. For the rhythmic aspect, capturing the cyclical nature of rhythms generates rhythms that are more attuned to the style of the music and enhances the rhythmic sense of the music. For harmony, a grasp of musical structure enables the selection of more appropriate chords, improves the fit with melody and rhythm, and enhances musical wholeness and harmony.

The experimental results show that compared with the traditional Transformer model, it performs better in indicators such as scale, pitch, and level, indicating that the quality of the generated music is higher and the generated music is more natural [7]. Ablation experiments verified the effectiveness of modules such as the Multi-Granular Attention Mechanism and the Genre-Assisted Categorization Discriminator, with metrics decreasing after removal of the Multi-Granular Attention Mechanism, demonstrating the importance of control over the learning of musical structures and genres.

2.2. VAE technology

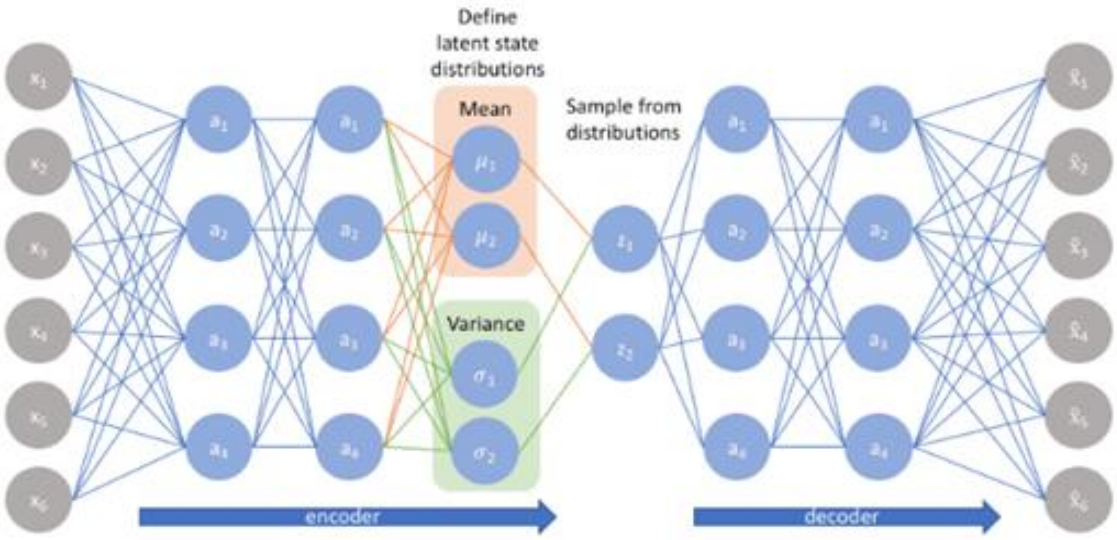

Figure 2. VAE structure diagram[8]

The principle of VAE lies in building an encoder-decoder structure that incorporates the idea of variational inference. The encoder is responsible for mapping the incoming music data (e.g., note information in MIDI-format music, including pitch, start time, duration, and volume) into a low-dimensional potential space. In this process, the encoder employs a bidirectional LSTM, which allows the model to capture both past and future information on the current note, leading to a comprehensive understanding of the music logic. For example, when processing a melody, it is possible to encode the current note based on the characteristics of the preceding and following notes. As shown in the fig. 2, the VAE maps the music data to the latent space by means of an encoder that defines a normal distribution in this space, and then reconstructs the music data from the latent vectors by means of a decoder. A probability distribution in the latent space, assuming a normal distribution N (0, I), is determined by calculating the mean vector μ and variance vector σ^2, implying that the model attempts to make the distribution of the generated latent vectors z close to the standard normal distribution, thus simulating the distribution of real music data in the latent space. And the task of the decoder is to reconstruct the music data from the potential vector z obtained by sampling from the potential space. With the help of a bidirectional LSTM network, the latent vectors are operated with the weight matrix W_x and the deviation vector b_x of the decoder network, and the reconstructed music data x' is obtained after the activation function.

In terms of musical structure, the encoder captures the dependencies between notes, which are preserved and reflected in the decoding to generate new music. For example, there is information about melodic development patterns, rhythmic repetition, etc., that is encoded into the latent space during the training process, helping to maintain the musical structure. In terms of music style, the learning process can be extracted from a large number of different styles of music data from the respective feature patterns, can be based on the location of the potential vectors to determine the style tendency, such as the potential vectors are in the region related to the style of popular music, the generated music is more capable of presenting the stylistic characteristics of popular music.

2.3. GAN and Music Generation

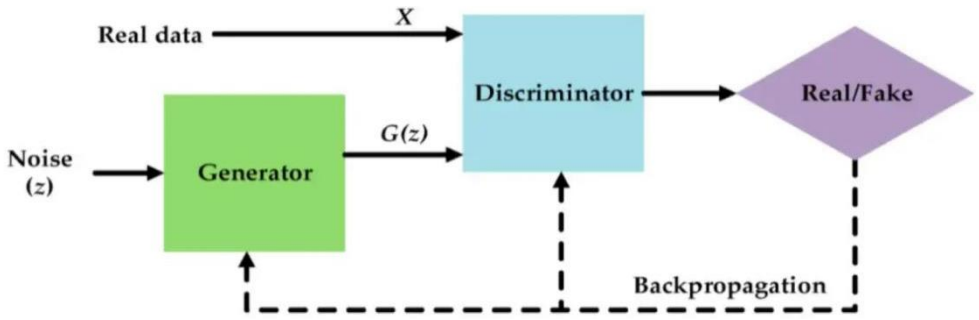

Figure 3. GAN structure diagram [9]

GAN mainly consists of two networks, the generator and the discriminator, which fight against each other, as shown in Fig. 3, the generation and quality of music is achieved through a process of confrontation. The generator generates music samples based on given conditions or random noise. In multi-track music generation, the generator generates melodies and rhythms that conform to the desired style, and the discriminator determines as accurately as possible whether the output is a real music sample or a generator-generated sample. After continuous training, the generator gradually learns the ability to generate high-quality music, and the discriminator continuously improves its discriminative ability.

In terms of generating multi-track music, taking the model proposed by Wang Tao et al. as an example [10], its architectural design characterizes the application of GAN in this field. Data-wise, the MIDI files are extracted from a variety of features (e.g., Bar, Position, etc.) and encoded into a sequence of events that provides the basis for the model to process the musical information. As a whole, three generators are included to learn the sequence information of a single track for piano, guitar, and bass respectively, encode the three tracks into a time series, and the generators capture the temporal structure of notes and feature relationships within a single track. At the same time, 6 CT-Transformer modules are used for pairwise learning of track sequences, e.g., when the piano track learns the information of the guitar track and the bass track, by defining the query, the key-value pairs and the Cross-Track attention mechanism, it realizes the interaction and fusion of the information of the different tracks, so as to make the generated music more harmonious. Finally, the discriminator discriminates between the real and generated sample sequences and guides the generator to improve the quality of the generated music.

In terms of style conditioning, GAN introduces style-related conditions in its generation, such as the rhythm, harmony, etc. of a particular music style, so that the generator generates the corresponding style of music according to different style requirements [11]. The discriminator, on the other hand, needs to judge whether the music conforms to the target style on the basis of the original, prompting the generator to learn the characteristics of different styles of music and realize style adjustment.

2.4. Other related technologies

Deep Reinforcement Learning in the automatic generation of harmony has been proposed for quantized encoding of chords and harmonies through the AHG-DRL algorithm [12]. The quantization of the chord is determined by the intervals of the two chord tones within the chord and the position

of the intervals, the lower the position of the intervals the greater the weight, and the root of the chord is the dominant and the value will be reduced, thus transforming the harmonic search space into a continuous space, so as to make the generated harmonies more varied. In addition, the deep reinforcement learning-based harmony generation framework contains a pre-training module, a harmony generation module, and a harmony refinement module. The pre-training module uses inverse reinforcement learning to pre-train the data in the harmony dataset, and constructs a reward function optimization strategy to equip the generated harmonies with basic musical functions. The harmony generation module takes emotion curves as input, the environment randomly generates curves in the training phase, and inputs the desired curves in the generation phase, and the model predicts the next chord based on the curves. The harmony refinement module modifies the generated harmonies to conform to the compositional specifications, checking the harmonic head modulation and tail termination, and adding missing elements if necessary.

The MICW model performs well in handling the task of multi-instrument music generation, and is able to accurately harmonize the performances between the different instruments so that the generated music is more harmonious [13]. For example, when generating multi-instrumental compositions, it is possible to take into account the relationship between the instruments and generate music that more closely resembles the effect of real compositions based on their characteristics. At the same time, the model is flexible use of compound words can produce more diverse musical fragments to meet the creative needs of different musical styles.

3. Literature application/experimental analysis

3.1. Movie Music Generation Based on Multi-Granularity Attention Transformer

This model brings new ideas to the creation of movie soundtracks. For example, experimenting on constructing a symbolic music dataset containing genre information, the model can generate matching music based on different genres according to the set target music style [7]. In terms of objective evaluation indexes, the model is closer to the real data than other mainstream style control models in terms of SC, PR and UPC, which indicates that the model, after learning the music data, can generate music that is more natural, with richer variations and characterized by the target genre, which enhances the synergistic effect between the music and the movie plot.

Further, the model can be adapted to the movie plot and emotional atmosphere. For example, if the movie proceeds to an important plot point, music with a tight tempo and tense melodic changes can be generated to enhance the atmosphere. Or scenes that need to express feelings can also generate melodic, slow-paced music.

In this regard, it can be seen that the model has many advantages over the traditional way of creating movie music, and the algorithm can shorten the creation cycle and meet the demand for large volumes. A genre-assisted classification discriminator can also be used to improve the classification probability of the genre to which the output belongs with higher accuracy than DeepJ and CPS. The generated music is also closer to real music in terms of subjective evaluation scoring. Nonetheless, the model can generate music based on the plot, but in the face of complex and varied segments with excessive emotional nuance, the flexibility and accuracy of the generated music is still insufficient compared to that of good composers, and the technique needs to be further optimized in this regard.

3.2. Experimental evaluation of the VAE music generation model

In VAE-based music generation models, the diversity of generated samples is an important criterion. Experiments were conducted to observe the effect on diversity by adjusting the parameter settings [14]. For the number of nerve cells in the hidden layer, the speed of convergence of the Loss value decreases as the number of nerve cells increases, indicating that the model generates a smaller error between the predicted value and the target value, and the model can learn more detailed information and generate more diverse music. For the number of training iterations, when the iteration to 2000 times, the generated spectrum has a lot of ups and downs, but there is a large gap between

the distribution and the sample spectrum, when the iteration reaches 4000 times the generated spectrum appears to be regular and similar to the sample spectral distribution, indicating that with the deepening of the training, the model learns the characteristics of the sample music, so as to generate the samples that are closer to the real music, and to increase the diversity of the samples.

3.3. Generative Adversarial Networks Results in Multitrack Music Generation

In multi-track music generation, GANs are able to effectively learn the interactions between different instruments, such as piano, guitar, and bass, by using the Cross-Track Transformer module to generate better coordinated music pieces. In the experiment, the music generated by the Transformer-GAN model was subjectively evaluated and the work was rated against the music created by professionals [10]. The results showed that the average score of human works was slightly higher among professional composers, but the average score of music created by the model was higher than that of human works among all participants, 8.11, indicating that the quality was very close to the level of the composers, generating very high quality works. In the comparison experiment with the other models (MTMG, HRNN, and MuseGAN), participants rated the music in four aspects, including melody, and Transformer-GAN scored significantly higher than the other three models in rhythm, melody, and harmony, suggesting that the structure of the generated music is more complex and conforms to the rules of music. Finally, prediction accuracy, sequence similarity and resting metrics were chosen to assess the similarity of the model-generated music to real music. The tracks generated by Transformer-GAN basically achieve good results, for example, the piano track improves 3% in real sequence similarity compared to the MuseGAN model, which indicates that the model has certain advantages in style control and music imitation generation, and can accurately simulate the note combinations and structural features. Although all the scores on the rest indicator are relatively low, it is also a side note to have a better generative effect, as the proper utilization of rests is crucial to the rhythm of a musical piece.

4. Suggestion

4.1. Challenge

4.1.1. Technological bottleneck

Current AI generative models face many challenges in music composition. Model training has a huge demand for computational resources, and training of complex neural network architectures requires high-performance devices and long hours of training. The long-term structural coherence of the generated music is difficult to guarantee, and the problem of reasonable localization but overall looseness can occur. The emotional expression of the generated music is not precise enough to convey subtle emotional changes.

4.1.2. The Dilemma of Integrating Art and Technology

How to incorporate artistic creativity and human emotion into the process of generating music is a major challenge. The generated music is slightly mechanical in listening, lacking the depth of emotion and unique creativity of human creation. Music is a kind of art, balancing technological innovation and the essence of musical art is very difficult, excessive pursuit of technical indicators is easy to ignore the connotation of art, affecting the value of the generated works.

4.2. Prospect

The future promises to improve existing techniques or the emergence of new ones. More efficient model architectures or reduced computational resource consumption. Intelligent training algorithms are expected to enhance the model learning ability, improve the quality of generated music while making breakthroughs in music understanding, generating diversity, accurately controlling stylistic characteristics, and generating more innovative and high-quality works. The application scenarios of artificial intelligence generative modeling in the field of music creation will continue to expand. Aid

students in music education by assisting them in their creative endeavors. Improve efficiency and quality in music production. There is also potential in some emerging fields, such as healthcare, where personalized music can be generated to assist treatment based on the mood of patient.

5. Conclusion

Artificial intelligence generative modeling has yielded significant results in the field of music composition. Multi-granularity Attention Transformer-based models improve the quality of movie music generation, VAE models perform uniquely in music structure and style learning, GAN has good results in multi-track music generation and style adjustment, and Deep Reinforcement Learning and MICW models also have their own roles to play. However, there are still many aspects that can be improved in this field. There are great prospects for technological progress and application expansion, and continuous innovation to promote the development of music creation.

References

[1] Zhang Z, Li L, Zhang J, Hu Z, Wang H, Yan C, Yang J, Qi Y, Generating High-quality Symbolic Music Using Fine-grained Discriminators, arXiv preprint, 2024.

[2] Souza G, Figueiredo F, Machado A, Guimarães D, do we need more complex representations for structure? A comparison of note duration representation for Music Transformers, arXiv preprint, 2024.

[3] Wang Y, Yang W, Dai Z, Zhang Y, Zhao K, Wang H, MeloTrans: A Text to Symbolic Music Generation Model Following Human Composition Habit, arXiv preprint, 2024.

[4] Kundu S, Singh S, Iwahori Y, Emotion-Guided Image to Music Generation, arXiv preprint, 2024.

[5] Epure E V, Meseguer Brocal G, Afchar D, Hennequin R, Harnessing High-Level Song Descriptors towards Natural Language-Based Music Recommendation, arXiv preprint, 2024.

[6] Transformer Model Explained, https://www.cnblogs.com/LXP-Never/p/15850142.html, 2024/11/28.

[7] Xiong X, Xie Z, Huang D, Zhu Y, Research on Film Music Generation Based on Multi-granularity Attention Transformer, Modern Film Technology, 2024, pp. 18 – 25.

[8] [VAE] Principles, https://www.cnblosg.com/yifanrensheng/p/13586468.html, 2024/11/28.

[9] Generative Adversarial Networks in Deep Learning in One Post | 8 Years of GANs Architecture Development, https://www/cnblogs.com/wxkang/p/17133320.html, 2024/11/28.

[10] Wang T, Jin C, Li X, Tie Y, Qi L, Multi-track Music Generation Adversarial Network Based on Transformer, Computer Applications, 2021, pp. 3585 – 3589.

[11] Wang W, Li J, Li Y, Xing, X, Generating Stylistically Adjustable Music Based on Transformer-GANs, Frontiers of Information Technology & Electronic Engineering, 2024, pp. 106 – 121.

[12] Liu Z, Liu W, Ran L, Xu K, Jiang Y, Li Q, Qiao S, An Automatic Harmony Generation Algorithm Based on Deep Reinforcement Learning, Radio Communications Technology, 2024, pp. 985 – 992.

[13] Liao Y, Yue W, Jian Y, Wang Z, Gao Y, Lu C, MICW: A Multi-instrument Music Generation Model Based on Improved Compound Words, In Proceedings of the 2022 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), 2022, pp. 1 – 10.

[14] Bai Y, A Music Generation Model Based on Variational Autoencoder, Computer Knowledge and Technology, 2024, pp. 15 – 18.