Proceedings of the 2023 International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/8/20230188

Deep learning-based music generation

Gening Lu

Zhuoyue Honors College, Hangzhou Dianzi University, Hangzhou, 310018, China

lugn@hdu.edu.cn

Abstract. Music is one of the greatest inventions in human history. Traditionally, music composition is time-consuming and complex, requiring master sound knowledge based on music theory and musical intuition. In recent decades, deep learning have been applied in music generation, and it has experienced the process from simple sequence generation to multi-track generation considering musicality, multiple methods are implied in study to generate better music and combine existing music theory with deep learning technology, while the current technology already allow people composite easily even without domain knowledge and massive manpower. This paper offers an overview of automatic music generation task, covering majority of the currently popular deep learning-based music generation models. In addition, in latter section discusses how to unify objective criteria for music on subjective way, proposes existing deficiencies and expands possible directions. The research in this review has significant foreshadowing meaning and reference value for developing music generation in the future.

Keywords: music generation, multi-track, deep learning.

1. Introduction

Music is closely related to human life and is widely used in music appreciation and multimedia audiovisual. However, with the development of society, people's demand for entertainment and emotion has increased, traditional artificial music creation could no longer satisfy the social demand for music. Traditional music creation requires sound knowledge in the field of music and certain musical instrument skills, even for professional composers, music composition is still intricate and time-consuming. How to use computer technology to assist or even replace human beings to create music has become an emerging research direction. Music Generation refers to computer composite automatically by applying algorithm and models. Differ from human composition, computer composition need to process music in forms that comprehensible to computer, turn music knowledge to effable constraints for the computer, to generate a piece of new music. Therefore, music generation is mainly focused on two parts: musical data representation and generation model construction.

From the traditional sequential modeling method at the beginning to the current generation technology based on a various models, the effect of music generation results has been tremendously improved, as the interest in automatic music composition never ceased. In 1956, the first computer-generated music composition Illiac Suite, a string quartet, was born by Lejaren Hiller. Inchoate experiments came up in the early 1980s such as the Experiments in Musical Intelligence (EMI) proposed by David Cope, who also put forward combine Markov chains with grammars to generate music automatically. However, music is consists of complex characteristics[1]. As for live or record audio

DOI: 10.54254/2755-2721/8/20230188

music, it contains emotional expression and interpretation of performers, and may show different characteristics from origin score by simply changing tempo. The early exploration of music generation mainly uses sequence modeling, while considering few about music theory and musicality, lack of generalizability capacity and rule-based definitions, lead to insufficient utilization of deeper music features. In recent decades, with the development of deep learning, many music generation tasks attempt to apply various Deep-Learning-based structure, which performed well in carrying out extensive tasks in image & video processing and natural language processing. The DeepBach model consisted by Long Short-Term Memory (LSTM), later the models named DeepJ and XiaoIce Band both made great success. As Transformer model emerges, MuseNet and Music Transformer made the results even natural and creative [2-5].

Admittedly, though music generation have made great achievements, several open questions still exist. Firstly, music generation combined with music domain knowledge, rather than just relying on machine learning methods for modeling and generation make result pieces more receptive, while how to turn theories into rational constraints is still under exploration. Particularly, the arrangement of chords and polyphony is a delicate art. Multiple key factors that related to music are worth exploring, such as musicality, mode, dynamics, chords, the structural features between movements, and the choice of timbre. Besides, aiming to fully show their charm to audiences, existing music pieces are multi-track to the majority, which call for arrangements. Usually there is a main melody track and the rest of the music tracks serve as accompaniment, they shall cooperate with each other tightly, otherwise the pieces would sound chaotic. Ensuring the harmony of multi-track music is an urgent problem to be solved. Moreover, for the sake of better meeting with the needs of modern music, the generated pieces are usually assigned with certain style requirement, which calls for style control. Nevertheless, the small corpus and lack of ideal data set makes training even difficult, causing generation models of specific musical genres with good performance have yet to appear. Therefore, how to generate music with corresponding style in the absence of parallel data of music genre is also a big issue. Researchers in today are trying hard to solve the above problems, developing new algorithms and innovate in models to achieve better results.

2. Forms of music data

Music could be stored in many forms, to further process music data, they shall be pre-processed into data types that be comprehensible for computer. On the same time, for different generation tasks should choose corresponding data set for training and test. Since the development of music generation, a series of music data sets emerged, covering classic, pop, jazz and other genres. The information varies with the degree to which the computer can edit it for different data types. As for digital music, it is generally editable for computer such as Music Instrument Digital Interface (MIDI) and piano roll, which contains music information about pitch, duration, timbre and dynamic of notes. Audio music, on the other hand, is not fully editable, it is about signal processing essentially, and could be converted into Mel-Spectrum then apply music transcribe technology to turn to MIDI and other formats.

3. Data set

Various music data sets have born since the development of music generation, different data set serve varied purposes and accommodate diverse tasks, mainly decided by the genre and quantity of music pieces it contained, and the selection of researchers is also strongly decided by the quality and format. Main formats of music data set for generation are stored in format of audio, MIDI, MusicXML etc. For the data-driven deep learning model, sufficient data support is necessary, followed by homogeneousness. An excellent model shall possess the quality of generalizability, easily distinguish the subcategories if the data set is heterogeneous. This made the data set selection almost as significant as how to build the model structure. Moreover, since music data sets are always expensive and may involve some copyright issues, plus scarce of digitalized data sets unlike other fields in deep learning, many commercial companies and research institutions are unwilling to make their own data set open source, the data set available for independent researcher and students in university are very limited actually. Table 1 summarized some available data sets categorized in storage formats.

DOI: 10.54254/2755-2721/8/20230188

3.1. MIDI

The most widely used music standard format in music domain, uses digital control signal to record notes and not transmit music signals, making it space-saving. It contains information about pitch, tempo, timbre, dynamic etc.

3.2. MusicXML

Based on digital score, it contains more messages than MIDI does. including music bar, slur, key and clef, note type and rest, the elements unable to precisely determined in listening, is also a general format for composing software.

3.3. Pianoroll

It represents each notes in 2D matrix, with using python package pipianoroll make it convenient to carry out reversible conversion between MIDI and pianoroll formats.

3.4. Text

Chris Walshaw developed ABC format, originally it was designed for folk music in West Europe, and was later extended to represent full classic music score.

3.5. Audio

The most common format, support playing directly on media player, the nature of audio is waveform, which could be stored in as wav, mp3, acc etc.

3.6. Multi-modality

Includes multi-mode information like score, lyrics and audio, mainly used for fusion generation tasks.

Table 1. Outline of various data sets.

Format	Name	Modality			Whether Polyphonic and More Description
		Score	Performance	Audio	
MIDI	JSB Chorus	√			Poly, 402 Bach four parts chorus.
	VGMIDI	√	√		Poly with sentiment, 823 piano video game soundtracks.
	Lakh MIDI	√	√		Multi-instrumental, 176581 files.
	Projective Orchestral	√			Orchestral, 392 MIDI files with piano and orchestral version.
	e-Piano Competition	√	√		Poly , ~1400 MIDI files of piano.
	BitMidi	√			Poly, 113244 files.
	Classical Archives	√			Poly, the biggest free classic music MIDI file dataset.
	The largest MIDI dataset on Internet	√			Poly & style, ~130000 in 8 genres.
	ADL Piano MIDI	√	√		Poly, 11086 piano MIDI files.
	GiantMIDI-Piano	√	√		Poly, 10854 files of piano pieces.
Music XML	TheoryTab	√			Poly, 16K lead sheet segments.
	Hooktheory	√			Poly, 11329 lead sheet segments.
	Wikifornia	√			Poly, 2252 westem music pieces.
	Musescore lead sheet	√	√		Performance, lead sheet of Yamaha e-Competition MIDI set.

DOI: 10.54254/2755-2721/8/20230188

Table 1. (continued).

Piano roll	Lakh Pianoroll	√	√		Multi-instrumental.
Text	Nottingham Music	√			Mono, ~1000 folk songs
	ABC tune book of Henrik Nobeck	√			Mono, >2800 ABC format, mainly Irish and Swiss traditional music.
	ABC version of FolkDB	√			Mono.
	KernScores	√			Poly, >700 million notes in 108703 files.
Audio	Nsynth			√	Music audio, 306043 notes.
	FMA			√	Music audio, 106574 tracks.
	Minist musical sound			√	Music audio, 50912 notes.
	GTZAN			√	Music audio, 1000 30s music segments.
	Studio On-Line (SOL)			√	Music audio, 120000 sounds.
	NUS Sung and Spoken Lyrics (NUS-48E) Corpus			√	Sing voice, 169 min recording of 48 English songs.
Multi-modality	MusicNet	√		√	Fusion, 330 recordings of classic.
	MASTERO	√	√	√	Fusion, 172 hours of virtuosic piano performances.
	NES Music Database	√		√	Multi-instrumental.
	Piano-Midi	√	√	√	Poly, 332 classic piano pieces.
	Groove MIDI	√	√	√	Drum, 13.6 hours recordings, 1150 MIDI files, >22000 measures of tempo-aligned expressive drumming.
	POP909	√	√	√	Poly, multiple version of 909 popular songs of piano .
	ASAP	√	√	√	Poly & fusion, 222 digital scores aligned with 1068 performances.
	Aligned lyrics-melody music dataset	√			Fusion, 1393720-notw sequences with 278740 syllable-note pairs.
	MTM Dataset	√		√	Fusion.

4. Automatic composition based on deep learning

Deep learning is a popular branch of machine learning in recent decades, as the artificial neural network research activated its original concept. It combines low-level features to form a more abstract high-level representation attribute category or feature, to discover distributed feature representation of data. The motivation of deep learning is to build a neural network simulating human brain for analysis and learning, which imitates human brain mechanism to interpret data. Compared with previous technologies, it particularly emphasizes the depth of model structure, has stronger learning and analysis ability of data characteristics, and is closer to the ultimate goal of artificial intelligence. Deep Learning-based music generation technology refers to that the ability to generate music automatically or semi-automatically. At present, there are two main music generation tasks, they are controllable music generation and automatic music generation. The former type of task requires certain input given by humans to assist composition, while the automatic music generation ab initio. This paper mainly discusses the automatic music generation.

Back to 1989, the first neural network model for music generation was the Recurrent Neural Networks (RNN), valid in learning sequence data. Nevertheless, RNN is always accompanied with gradient vanishing and error propagation problems, causing this model weak in store long history

DOI: 10.54254/2755-2721/8/20230188

information of sequences. Therefore, many researchers add LSTM mechanism to the RNN to assist network memory and sequence information retrieval. With the iterative development of deep learning technology, more deep learning structures are proposed and applied to music generation, such as Transformer, Generative Adversarial Networks (GANs) and Variational Automatic Encoder (VAE). These models are initially applied to image & video processing and generation tasks, is proved to perform well in these fields. Since music generation is much more complex than 2D image generation, and the nature of music generation in symbolic domain is discrete sequence generation, researchers have made great efforts to transplant and improve these model structure, attempt to make them more adaptable and exquisite. In following subsections, this review paper discusses the implementation details and existing deficiency of a-state-of-the-art models of recent years in each stage of automatic music generation.

4.1. Melody generation

Melody generation is the most fundamental core-part of automatic music generation. Melody refers to an organized rhythmic sequence of musical sounds formed by artistic conception, a logical monophonic component of certain pitch, duration, and dynamics. Hence, the essence of automatic melody generation is to predict the information about next single note or rest. The development of melody generation model has gone through the process from short sequence to long sequence. Part of the most famous melody generation models are shown in Figure 1 in form of chronology, collected by S. Ji et al. [6].

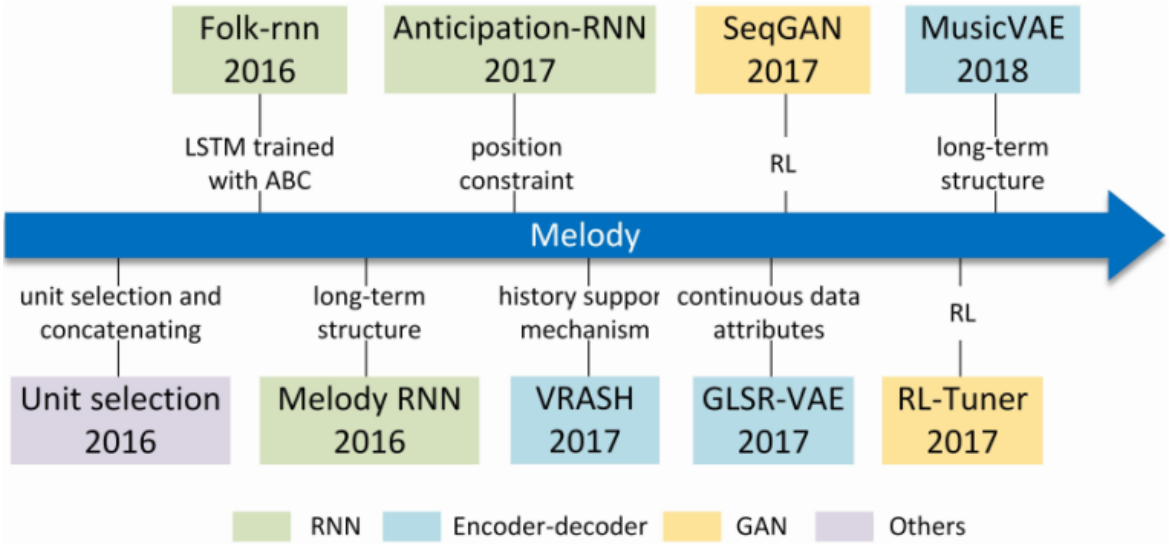

Figure 1. S. Ji et al. [5] collected this representative evolutionary chronology (2012 ~ 2020) of melody generation model, with color to distinguish the model or algorithm applied.

In 2015, Nayebi et al. [7] did channel merging and dimensionality reduction for data in waveform format by segmentation the time domain, then converted Discrete Fourier Transform (DFT) to frequency domain analysis, and compared LSTM with Gated Recurrent Unit (GRU), the results showed that the former one, LSTM do better than GRU in melody generation based on very small data set. 2016 was a big year for the development of melody generation, the first deep learning model for generate short melodies with RNN and semantic models such as Unit Selection was proposed by Bretan et al. [8]. In the same year, the Magenta team in Google published the famous Melody RNN model [9], which performs better than previous models in learning long-term structures. More than one baseline RNN model, it open-sourced two new Magenta models, Lookback RNN and Attention RNN creatively, allowing people to add more controllable information and enhance the interactivity with composers.

Besides the RNN model, VAE, Transformer and GAN etc. models are always combined with other generative models for melody generation. In 2018, Roberts et al. put forward a hierarchical latent vector model named MusicVAE, encourage the model to exploit latent variable coding to solve the common

DOI: 10.54254/2755-2721/8/20230188

problem of posterior collapse in VAE, support generating music by interpolating in a latent space, and this model could even generate chord for bass and drums [10].

To generate more inspiring long-term sequences of music with good sense of musicality and harmony, new models based on Transformers or GANs have emerged, some of their hybrid models with traditional deep learning models such as TransformerVAE by J. Jiang et al. in 2020, it performs well in generating chord music and support generating music phrases [11].

Particularly, the Natural Language Processing (NLP) models are already applied to lyric generation task. In fact, both psychological and anthropological studies have shown a correlation between the ability of human language and music, regarding neural representation and emotional recognition mode. NLP has been relatively mature in the field of speech recognition and language generation, while the music score generation in automatic music generation task can exactly correspond to the language generation in NLP field, which are all represented by symbolic tokens, so that researchers can learn from the deep learning technology in these mature fields to help music generation. Similarly, the single-track melody generation could be considered as special monophonic music, corresponding to the linear generation problem to some extent. Thus leading researchers to attempt to explore if they could further construct music melody combine with NLP methods by using music grammar, just as Perchy et al. [12] have proposed in their work that is based on stochastic text-free grammars. The field of melody generation is the foundation of various related tasks, especially accompany-generative task that need inputs, the result of melody generation task could be applied as input of other model and thereby reduce the dependency of searching extra data set for input and avoid self-similarity. Besides, melody generation task is often the first step in hierarchical study of complex automatic composition task corresponding to submodels, it plays an significant role in founding the research of music generation and even computer music.

4.2. Structure control

Music structure is of great weight in music theory. During classical music era, there are strict requirements for the structure of different music forms. Taking classical music as example, it is a kind of music with the nature of regularity, has the characteristics of balance and clarity, have various branches such as symphony, sonata, concerto, opera and chamber music etc., for each music form the requirements varied. In the classical period, composers generally followed the form and principle of consensus composition. For example, symphonies were generally composed of four movements, and sonata form included exposition, development, and augmentation, etc., with certain logical correlation between movements and phrases. A good musical structure is closely related to the melodic logic, music form conventions, and chord progression, the exact factor that makes structural control one of the most difficult tasks in music generation field.

Researchers attempt to imitate the musical structure by enforcing the self-similarity constraints on the model. In 2018, Lattner et al. [13] proposed Convolutional Restricted Boltzmann Machines Model, is combined with gradient descent constraint optimisation to control a high-level self-similarity polyphonic music structure in tonal etc. properties while preserving local musical coherence. Nevertheless, the generation model need to be realized by manually adding structural constraints for simulation. As for music generation task with less stringent genre structure requirements, Shuqi et al. [14] introduced a hierarchical music structure representation and a multi-step generative process, assembly called as MusicFrameworks, aiming to control the overall structure of the music, including chord, melodic contour and rhythm, thereby enable the result pieces to maintain the properties of paragraph logic and musicality completion. It has good controllability and interactivity, which is the innovate point superior to other models. The manual evaluation indicated that over 50 percent result pieces are considered better than the work of human being. However, up to now there is seldom exist music generation model that can automatically generate reasonable structured results under strict genre constraint, that is, the model independent from templates or high-level structural information passed to the neural networks.

DOI: 10.54254/2755-2721/8/20230188

4.3. Monophony, polyphony and chord feature

Chords and polyphonic textures play important roles in improving music aural sensation. Chord refers to a group of sounds with a certain interval relationship, that is, three or more notes, according to the overlapping relationship of the third or non-third interval relationship, combined vertically, there are generally 14 kinds of chords in music theory, and chords are partly depend on the tune of music, which means tonality. Set aside atonal music, tonality is the general name of the tonic and mode categories of the tune. For instance, the major mode with C as the tonic is "C major", and the minor mode with a as the tonic is "a minor", etc. By the same token, there are 24 major tunes in general music.

Monophonic music is a kind of multi-part music. The whole work is mainly carried out by the melody of one part, while the other parts are accompanied and accompanied by harmony or rhythm. Harmony is the most important creation basis of monophony, even till today the majority of modern popular songs are monophonic music. Polyphonic music is also a kind of multi-part music, it is a combination of rational rigor and emotional harmony, which tests the ability in composing music. Differed from monophonic music, the polyphonic music is characterized as including more than two independent melodies, the melodies are combined harmoniously through technical processing, hence, it could also be considered as the advanced version of single-melody generation task.

The research of monophony is closer to arranging accompaniment part for single melody, in following phrase this paper talks about chord and polyphony music generation tasks. Current researchers mainly focus on generating chord for melodies with limited chord types or based on probabilistic distribution model to improve the aural richness of music. The first work in polyphonic music is RNN-RBM model proposed by N. Boulanger-Lewandowski et al. [15], they put forwarded a probabilistic model based on distribution estimators and RNN, being able to detect temporal dependencies in high-level-dimensional sequences. DeepBach is an excellent chord generative model especially in generating polyphonic chant in style of Bach, proposed by G. Hadjeres et al. [2]. It was trained on Bach's works, among the greatest classic works of Baroque period, and is still regarded as the most rational music in the world, with a rigorous logical and harmonic structure. The DeepBach model is based on dependent network, produce Bach's chorale style works by Gibbs sampling, rather than the left-to-right generation pattern found in common generation models, able to generate very realistic Bach-style works, is of great significance to explore how to generate polyphonic choral music in more extensive styles. PianoTree is the newest model in polyphony music generation proposed by Z. Wang et al. [16] in 2020. With combining counterpoint technique in traditional composition theory with VAE model. PianoTree applied a novel tree-structure extension with VAE to adapt the polyphonic music learning, demonstrating the semantic meaning of latent code for polyphonic music and the validity of PianoTree model in downstream music generation. The developmental chronology of polyphonic music model and their characteristics are summarized in Figure 2.

DOI: 10.54254/2755-2721/8/20230188

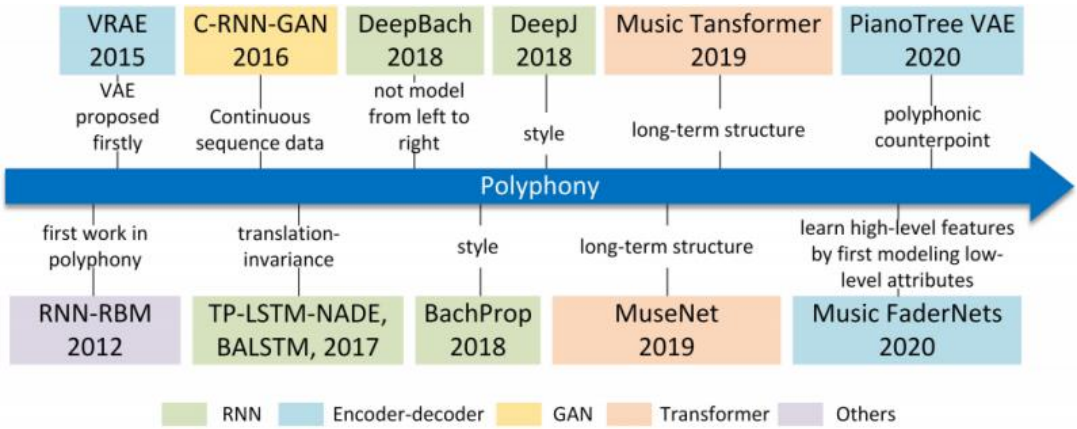

Figure 2. The representative evolutionary chronology (2012 ~ 2020) of main polyphonic music generative model with summarized features or meanings.

Polyphony music generation and chord generation not only It not only increases the richness of music hierarchy, but also facilitates the further industrialization of automatic music generation. The ability to quickly achieve a complete prototype of music, even if it is not processed in detail, is an important feature in marketization. The current challenge of the polyphony and chord generation task mainly lies in generate controllable quality music with stability.

4.4. Style transfer

Music style refers to a complex combination of characteristics ranging from music theory to sound design, such as e.g. classic, pop, rock, jazz etc. These features include the instruments selected in the piece, timbre, or the after processing effect in composition and arrangement etc. Other factors also matters such as data set for training since deep learning is characterized as data-driven model, the style of pieces generated by the ready-made application largely depends on the choice of data set, this feature indicates that music generation model of given style might be portable in certain cases. It is difficult for some music style generation models to be fully explored and developed at the present stage due to the unbalanced data, model training could unable to be carried out since scarce of appropriate or sufficient data and data set. One feasible valid solving method for this situation is using style transfer technology, inspired by achievements of style transfer application in image processing, learning features vectors of required style or using style embedding technique to control the style generated by adding notes or modify pitch interval etc.

The first successful example of symbolic music style transfer application emerged in 2018, MIDI-VAE model is proposed by G. Brunner et al. [17], realizing music style transfer automatically is performed by changing pitch and dynamics of notes as well as timbre, or interpolate short pieces between music phrases to create hybrid music, the available style ranging from classic to jazz, and carry out evaluation by training separate style validation classifiers for efficacy. Up to now, there are few effective models for music style transfer field, but it is a field with great commercial and research application potential. In addition to direct style transfer application, new music styles can also be generated by parameter adjustment, and even help developing brand new restylized data sets based on existing data sets for further learning and training tasks.

4.5. Multi-track arrangement

Multi-track arrangement could be characterized as both monophonic and polyphonic music. Assuming in arrangement phase, each timbre (part) occupies one sound track, the multi-track music is carried out with several timbre collaborating mutually, the sonic texture then create one or more melodies simultaneously. Due to its complex mechanism in numerous permutation and combination of timbres, and further considering that the interplay of parts, timbre and melody may full of randomness, it presents higher challenges to music generation tasks comparing to single instrument music whether it is

DOI: 10.54254/2755-2721/8/20230188

monophony or polyphony. In account of the intricate mechanism of multi-track music, research on this field has began to emerge in recent years, the evolutionary chronology of multi-track music generative model is shown as Figure 3.

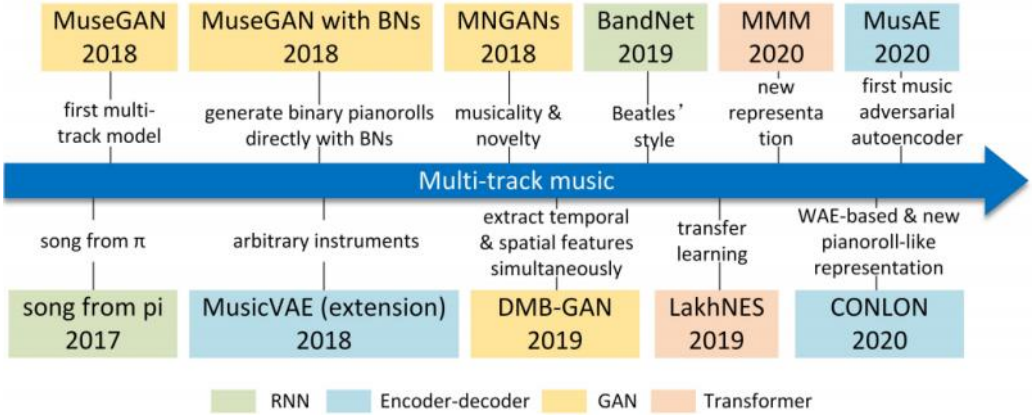

Figure 3. The representative evolutionary chronology (2012 ~ 2020) of main multi-track music generative model with summarized features or meanings.

As the first multi-track generative model emerged, the MuseGAN model proposed by Dong et al. [18], opened the gate of multi-track arrangement with deep learning technology. MuseGAN drew great emphasis on construct time-sequence model, thereby creating three models based on GANs separately under different assumptions: the jamming model, the composer model and the hybrid model. MuseGAN supports generate five-track music consist of string, piano, guitar, bass adn drums in format of pianoroll, the evaluation of results is controlled by a few of inter-track and intra-track metrics to measure if the tracks have good sense of alignment, and a novel point of MuseGAN is the ability to arrange creatively extra four-track to accompany the given single track (the input track is regarded as melody), made great encouragement for study in multi-track arrangement generation field. The first music adversarial auto-encoder model arose in 2020, MusAE model is presented by Valenti et al. [19], this model demonstrates that the Gaussian mixture with music information involved is able to be sufficient prior of the latent space of the autoencoder, enhances the reconstruction accuracy comparing to standard autoencoders. One pioneer point of MusAE is its capacity to generate vivid interpolations into musical phrase sequences and smoothly change the dynamic property of each track. In 2021, Jin et al. presented a new multi-track music generation model based on transformers named MTT-GAN [20], characterized as using the decoding block of transformer to learn internal information of the given single track input while utilizing cross-track transformers to learn the information of inter-track and intra-track between various musical instruments. And most recently, Jiafeng et al. [21] published a paper about creating a model called SymphonyNet based on permutation invariant language model and have achieved very soul-stirring results in generating multi-track magnificent symphony score. In order to maintain the local structure of music in multi-tracks and fully consider the features of semi-permutation invariance in symbolic music domain, the authors proposed the Multi-track Multi-instrument Repeatable (MMR) representation with special 3D positional embedding. the quality of finished pieces are even suffice to support film soundtrack marvelously, marking a new monument to the symbolic domain multi-track music generation. In sum, multi-track is a field with promising future in aesthetic and industrial application, and the state-of-the-art pioneer model already achieved great feats, and the current challenge is mainly lies on generate staple high-quality music while exploit more opportunities.

5. Evaluate methods

Music has developed from classical period till current days, there is still no unified evaluation standard, while it is of necessity to evaluate the generated music pieces scientifically and reasonably, which help to analysis and thereby further improve the model. Considering the great influence of the subjectivity of listening sense on the evaluation, it is difficult to establish a reasonable evaluation system for generating

DOI: 10.54254/2755-2721/8/20230188

music model. In the music industry, most of the comments on music are based on the subjective feeling of audience, and the relatively authoritative scientific critical analysis is often given by music experts and people in related industries or at least interested in music. At the same time, one thing has to be admitted is that different music performance may also cause great discrepancies in aural sensation due to e.g. the huge effect of playing skill pose to the effect of music that the music characteristic performers carry out. Even for master violinists, they interpret the same piece of Mozart's concerto in various style as "the second composer", the audience as "the third composer" may still tend to hold their own preferences for that. Therefore, it must be admitted that the subjectivity exists objectively in the evaluation process. A reasonable evaluation method should take into account both subjective and objective indicators, and dynamically adjust it according to the music forms and types generated by the model. In addition, the evaluation method for generation result without timbre processing (e.g. MIDI, MusicXML, Pianoroll etc.) shall vary from the method apply to evaluate generated audio performance, which contains much more sonic details and variations than the former class. The following part shows several feasible methods for manual and objective evaluation with indicators for generated music in symbolic domain.

5.1. Manual evaluation

A common approach is to invite a sufficient number of third-party professional musicians to participate in the testing and evaluation of the generated music, such as DeepBach project [2] had invited more than 1600 people including 400 musicians and music major students to recognize the generated results from the mixture set of Bach and DeepBach. The manual evaluation mainly focuses on perceptual cognition of music, including rhythmic aesthetics, completeness, harmony, stylistic rationality, naturalness, emotional expression and overall effect etc.. Due to strong subjectivity, perceptual indicators are difficult to quantify, which can be compensated by a large number of scoring data, invite testers from diacritic background in music (e.g. professors with sound theoretical music knowledge, experienced musicians, music major students, and amateur of music etc.), and carry out hierarchically analysis on different indicators, so as to balance the problem of indicator-imbalance. Thereby the evaluation result is able to reflect whether the generated music conforms to people's tastes to some extent.

5.2. Objective evaluation

Unlike the manual evaluation that generally based on perception recognition and auditory sense, the objective evaluation tend to carry out characteristics numerical analysis for music in symbolic domain, and have dependence on the music format being evaluated according to the recording information involved in each note, e.g. comparing to MIDI, pianoroll format is scarce of dynamic information. Part of available indicators and their definition are shown in Table 2.

Table 2. Available indicators.

Indicator	Definition
Loss	A common indicator in measure the variance between input and output
Perplexity	A common indicator in measure the generalization of model
Rest bar rate	The rate of rest bar
Pitch range	The range of the note in single bar
Tonal Distance	Equal to tonic distance, ain to measure the harmony of chord
Pattern Repetition	To measure the structure of music by detect melody repetition

In addition to harmony analysis by manual perceptual recognition, a function to assess music harmony could be described as chord similarity since the tracks in a harmony music have similar chord component, thereby define the harmony score as:

DOI: 10.54254/2755-2721/8/20230188

$$Harmony\ Score = \sum_{n=1}^{N} \omega\left(\bigcap_{k=1}^{K} C_n^k\right) \tag{1}$$

The function $\omega(x)$ is defined as:

$$\omega(x) = \begin{cases} 1, & if\ x \neq \emptyset \\ 0, & if\ x = \emptyset \end{cases} \tag{2}$$

And similarly, other factors that could only be measured by manual perception with intuition could follow the example of harmony score, thereby able to be evaluated in quantifiable value, gradually realizing the goal of quantifying perceptual indicators numerically and enhancing the generalization ability of inter-model result comparison.

6. Discussions

According to the above studies, automatic music generation modeling requires profound insight and in-depth comprehension of music. The existing music generation models based on deep learning have achieved brilliant achievements, but still face with various challenges. In addition, deep learning, like machine learning, is based on the logical processing of abstract features, which is generally not interpretable in specific problems. This section focuses on the challenges that music generation faces in current development and suggests possible future research directions.

6.1. Emotion

The emotional tension expressed by music is the soul of it and the part of most charming, arise moving sensation inside audience. Current generated music is partially able to control their harmony and structure, but it still remains a mystery to researchers for how to generate music in particular emotion and be dulcet with musicality. Assuming the training part need to classify the data into sections according to the emotion tag could help carrying out music emotional recognition mechanism and thereby further expand music property perception.

6.2. Creativity

Some music sounds pleasing due to their familiarity to ears, while music composition is a complex mixture of creativity and also follow rules such as musical theory. Since deep learning models are data-driven, the generated music distribution is often highly similar to the adopted data set. One idea that might work is create interpolation and insert them into original generated music sequence. If therefore go into infinite iteration, researchers shall consider how to change the model so that it can eventually converge and end the iteration.Although the DeepJ [3] model is able to generate creative music in style of classic, it is still immature to explore more styles and the boundary of creativity. Currently, it might work in generating short sequence and insert interpolation, but the biggest question is guarantee the quality of generated pieces with interpolation in a long sequence.

6.3. Data absence

Training deep learning-based models requires massive amounts of data, especially the Transformer model, which performs relatively well so far. However, due to the relatively sparse data set, and serious non-parallel style problems of data set, the model unable to be sufficiently trained due to the limitations of the original data set during training. The choice of research direction and style of generated pieces is also greatly influenced and limited. One solution, to be explored, is to perform transfer learning via pre-trained models. And as mentioned in 4.4, when the style transfer technology becomes mature enough, they could be applied to create new data sets. In addition, the number of instruments in the data set is also extremely unbalanced. Instruments with uniform pitch, such as piano, guitar, and drum, have relatively rich data, while string instruments with more elaborate and complex sound principles lack data that can fully express spectral details. Further, as mentioned in the Introduction, another problem

DOI: 10.54254/2755-2721/8/20230188

about data set is that the music data set is expensive and may be involve in copyright issues, driving researchers prefer to train on classical music. A large number of unpublished data sets and the high computational costs have also raised the threshold of research in this field, causing that existing automatic music generation models mostly focus on single style and this may weaken the research toward the model ability of generalization.

6.4. Variable and controllable length music sequence

Some existing models can only generate fixed-length segments, or the generated music length is controllable but be abrupt in the end of phrases owning to Blunt time step control. The focus is to improve the quality of music closure, which is one of the key points of generating Turing test for human-computer music.

6.5. Interactivity

The general lack of interactivity in current music generation models prevents them from being readily usable as auxiliary composition tools, specific performance as weakness on adding more flexible and complex constraints to the model by users. The existing regulation methods are not aimed at the overall control of the generation goal but at local unary constraints such as pitch, rest, tempo etc., adding to the fact that the output of deep learning model is inexplicable, make it difficult for users to both fine control and coarse control their output to approach their target requirement in a easy way, as the core problem is that most models and deep learning methods are week in music fine perception and property control.

6.6. Complex grace notes

In reality, traditional composers often add many ornamental grace notes to their music in addition to performance skills to improve the musical dynamic and enhance contagious expression. However, this is difficult to be reflected in the existing data set, and the data-driven music generation model often unable to capture these special fine features. For example, the boeing, vibrato which is common in violin performance may be recorded as a single tonic with a very short duration, glissando may be smoothed by MIDI, the timbre of overtones differs from that of normal performance, etc. These feature capture of grace notes is unlikely to be achieved at present, since even experienced players are sometimes confused about the duration and strength that grace notes should occupy. In general, the theoretical research on this area is still in a stage of relying on perceptual knowledge, and the closest first step is to introduce forms that can express richer details of musical notation in the future.

6.7. Interpretability

One of the most important ethical issues plaguing artificial intelligence researchers is the feature of "black box" of AI, that is, the result and processes of AI model is generally uninterpretable. E. R. Miranda et al. [22] are making great efforts to develop a new intelligent music system based on Quantum Natural Language Processing (QNLP) that is interpretable compositional, aiming to eliminate the shortness of uninterpretability involved in artificial intelligence models. The core model, Distributional Compositional Categorical (DisCoCat) is a natural framework to develop natural language-like generative music systems that apply to quantum computing. They even developed a test system called Quanthoven to examine the performance of quantum classifiers on the perception and controllability of musical compositions, uses generation-and-test approach developed by Lejaren Hiller, the pioneer in using algorithms to create computer music, could be further combined with evolutionary computation techniques for music composition (e.g., genetic algorithm, heuristic algorithm). At present, there is no system that can fully explain the "black-box" nature of artificial intelligence models, but the work of E. R. Miranda et al. is a beneficial attempt, it means more than what it is in automatic music composition field. Some research indicates that cognitive psychology may be needed to solve this problem for good in future.

DOI: 10.54254/2755-2721/8/20230188

7. Conclusions

This paper describes the origin and latest developments in each aspect of automatic music generation task based on deep learning hierarchically, from perspective of divided stages of automatic music composition: melody generation, structure control, polyphony and chord generation, style transfer and multi-track arrangement. This paper also discusses the difficulties and challenges faced by the automatic music generation tasks based on deep learning at present, puts forward some possible solutions, briefly describes the generally used data sets and common music storage format as well as representation methods in existing research. Evaluation methods based on both subjective and objective aspects are also introduced, so as to some possible future development directions, like style transfer, generate music with assigned emotion and style, introduce model with more interactivity that support more flexibility on devising generated results etc. In sum, the past decade of research on automatic music generation has yielded promising results, though problems in terms of structure, harmony, creativity, interactivity, and detail diversity still exist and may unable to be solved completely in recent future, is still far from the ultimate expectations. Neither the quality of generation nor the control of detail can replace human workers yet, but it is worth to be expected that as the technology continues to develop, people will eventually acclimatize the way that carried out the relationship between human and AI composers, and draw a blueprint that make use of their very best qualities.

References

[1] D. Cope, "Experiments in musical intelligence (EMI): Non-linear linguistic-based composition," Interface, vol. 18, no. 1–2, pp. 117–139, Jan. 1989, doi: 10.1080/09298218908570541.

[2] G. Hadjeres, F. Pachet, and F. Nielsen, "DeepBach: a Steerable Model for Bach Chorales Generation," in 2017 ICML 34th International Conference on Machine Learning (ICML). ICML, 2018, pp.1362-1371. doi: 10.48550/arXiv.1612.01010.

[3] H. H. Mao, T. Shin, and G. Cottrell, "DeepJ: Style-Specific Music Generation," in 2018 IEEE 12th International Conference on Semantic Computing (ICSC), Laguna Hills, CA, USA, Jan. 2018, pp. 377–382. doi: 10.1109/ICSC.2018.00077.

[4] G. Barina, A. Topirceanu, and M. Udrescu, "MuSeNet: Natural patterns in the music artists industry," in 2014 IEEE 9th IEEE International Symposium on Applied Computational Intelligence and Informatics (SACI), Timisoara, Romania, May 2014, pp. 317–322. doi: 10.1109/SACI.2014.6840084.

[5] C.-Z. A. Huang et al., "Music Transformer: Generating Music with Long-Term Structure." arXiv, Dec. 12, 2018. Accessed: Feb. 07, 2023. [Online]. Available: http://arxiv.org/abs/1809.04281

[6] S. Ji, J. Luo, and X. Yang, "A Comprehensive Survey on Deep Music Generation: Multi-level Representations, Algorithms, Evaluations, and Future Directions," J. ACM, Nov. 2020, [Online]. Available: http://arxiv.org/abs/2011.06801.

[7] A. Nayebi and M. Vitelli, "GRUV: Algorithmic Music Generation using Recurrent Neural Networks." 2015. [Online]. Available: http://cs224d.stanford.edu/reports/NayebiAran.pdf.

[8] M. Bretan, G. Weinberg, and L. Heck, "A Unit Selection Methodology for Music Generation Using Deep Neural Networks." arXiv, Dec. 12, 2016. Accessed: Feb. 05, 2023. [Online]. Available: http://arxiv.org/abs/1612.03789.

[9] E. Waite, "Generating long-term structure in songs and stories." Web blog post. Magenta, 15 (4), [Online] Available: https://magenta.tensorflow.org/2016/07/15/lookback-rnn-attention-rnn, 2016

[10] A. Roberts, J. Engel, C. Raffel, C. Hawthrone, and D. Eck, "A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music." arXiv, Nov. 11, 2019. Accessed: Jan. 31, 2023. [Online]. Available: http://arxiv.org/abs/1803.05428.

[11] J. Jiang, G. G. Xia, D. B. Carlton C. N. Anderson, and R. H. Miyakawa, "Transformer VAE: A Hierarchical Model for Structure-Aware and Interpretable Music Representation Learning," in ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal

DOI: 10.54254/2755-2721/8/20230188

Processing (ICASSP), Barcelona, Spain, May 2020, pp. 516–520. doi: 10.1109/ICASSP40776.2020.9054554.

[12] P. Salim, Gerardo M, and Sarria M., "Musical Composition with Stochastic Context-Free Grammars," presented at the In Proceedings of 8th Mexican International Conference on Artificial Intelligence, 2016. [Online]. Available: https://hal. inria.fr/hal-01257155. Accessed on 05 April 2021.

[13] S. Lattner, M. Grachten, and G. Widmer, "Imposing Higher-Level Structure in Polyphonic Music Generation Using Convolutional Restricted Boltzmann Machines and Constraints," Journal of Creative Music Systems, vol. 2, no. 2, Mar. 2018, doi: 10.5920/jcms.2018.01.

[14] D. Shuqi, J. Zeyu, C. Gomes, and R. B. Dannenberg, "Controllable deep melody generation via hierarchical music structure representation." arXiv, Sep. 01, 2021. Accessed: Feb. 24, 2023. [Online]. Available: http://arxiv.org/abs/2109.00663.

[15] N. Boulanger-Lewandowski, Y. Bengio, and P. Vincent, "Modeling Temporal Dependencies in High-Dimensional Sequences: Application to Polyphonic Music Generation and Transcription." in 2012 ICML 29th International Conference on Machine Learning (ICML). Jun. 2012, doi: 10.1002/chem.201102611.

[16] Z. Wang, Y. Zhang, Y. Zhang, J. Jiang, R. Yang and J. Zhao et al., "PIANOTREE VAE: Structured Representation Learning for Polyphonic Music." arXiv, Aug. 17, 2020. Accessed: Feb. 06, 2023. [Online]. Available: http://arxiv.org/abs/2008.07118.

[17] G. Brunner, A. Konrad, Y. Wang, and R. Wattenhofer, "MIDI-VAE: Modeling Dynamics and Instrumentation of Music with Applications to Style Transfer." arXiv, Sep. 20, 2018. Accessed: Feb. 07, 2023. [Online]. Available: http://arxiv.org/abs/1809.07600.

[18] H. W. Dong, W. Y. Hsiao, L. C. Yang, and Y. H. Yang, "MuseGAN: Multi-track Sequential Generative Adversarial Networks for Symbolic Music Generation and Accompaniment." arXiv, Nov. 24, 2017. Accessed: Jan. 31, 2023. [Online]. Available: http://arxiv.org/abs/1709.06298.

[19] A. Valenti, A. Carta, and D. Bacciu, "Learning Style-Aware Symbolic Music Representations by Adversarial Autoencoders," in 2020 ECAI 24th European Conference on Artificial Intelligence (ECAI), Feb. 2020. doi: 10.48550/arXiv.2001.05494.

[20] C. Jin et al., "A transformer generative adversarial network for multi-track music generation," CAAI Trans on Intel Tech, vol. 7, no. 3, pp. 369–380, Sep. 2022, doi: 10.1049/cit2.12065.

[21] L. Jiafeng et al., "Symphony Generation with Permutation Invariant Language Model." arXiv, Sep. 16, 2022. doi: 10.48550/arXiv.2205.05448.

[22] E. R. Miranda, R. Yeung, A. Pearson, K. Meichanetzidis, and B. Coecke, "A Quantum Natural Language Processing Approach to Musical Intelligence." arXiv, Dec. 09, 2021. doi: 10.48550/arXiv.2111.06741.