AI Music Therapist: A Study on Generating Specific Therapeutic Music based on Deep Generative Adversarial Network Approach

Yurui Hou
Walnut Hill School for The Arts
Natick, MA 01760, United States
yurui.hou2023@walnuthillarts.org

2022 IEEE 2nd International Conference on Electronic Technology, Communication and Information (ICETCI) | 978-1-7281-8115-8/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICETCI55101.2022.9832398

Abstract—**Music therapy models have been widely known in the medical field for their effectiveness in early childhood trauma, depression, and Parkinson's disease, and have been incorporated in the treatments of many medically advanced countries. The main musical tools in music therapy are listening and improvisation, but given that music composition requires the accumulation of much experience and time spent on arranging, neural networks are beginning to take effect in solving this problem. Adversarial Generative Network (GAN), Long-Short Time Memory (LSTM), and other neural networks learn data from prepared MIDI (Musical Instrument Digital Interface) files with therapeutic effects. Moreover, they generate recommendations of new therapeutic music of specific genres for users, helping them to perceive the style and structure of the new music. The experiments demonstrate that the designed GAN neural network makes the music more natural and smooth.**

Keywords—Music therapy, Music analysis, LSTM, GAN, CNN

I. Introduction

A. Origin of Music Therapy

The concept of music therapy was originated in the United States in the 1940s. It is said that during World War II, when American wounded soldiers had a high rate of infection and death, and were mentally depressed, a physician tried to play familiar songs from their homeland, and found that the soldiers' moods quickly stabilized, and even the healing period after surgery was shortened. As a result, a group of musicians became involved in hospital care after World War II.

Music therapy is actually a special type of psychotherapy in which the music therapist is specially trained to help and motivate the client to get a lot of mental help in music by providing a rhythmic or tonal foundation. Music therapy models include playing composed music on instruments, singing, writing or improvising songs, and listening to music.

B. Research Progress of Music Therapy

Past clinical reports and studies have reported that music therapy has contributed to some psychotherapy. Some music therapy models speculate that music may be helpful in addressing early childhood trauma. In addition, many studies and clinical evidence have demonstrated the effectiveness of music therapy for depression and Parkinson's disease. Many medical powers such as the United States, China, and Japan have now incorporated music therapy into their treatments. In 2000, Gold et al. conducted a study that revealed that adding music therapy to standard treatment had strong and significant effects on overall status, general symptoms, negative symptoms, depression, anxiety, functioning, and musical engagement. The findings suggest that music therapy is an effective treatment for patients with severe psychiatric and non-psychiatric disorders to improve their overall status, symptoms and functioning.

Music therapy not only has considerable efficacy in psychological disorders, but also has complementary effects in physical disorders. Wang Yuhua et al. played pop songs or light music to conscious patients during surgery, and observed the changes of their heart rate and blood pressure, and scored their anxiety. They found that patients had a significant reduction in systolic blood pressure and heart rate and a reduction in anxiety after listening to music for 30 min. It was concluded that the application of background music in the operating room is one of the effective measures to reduce the psychological stress of surgical patients.

C. Research Plan and Purpose of the Study

In the field of artificial intelligence, it has been regarded as one of the most exciting tasks to produce realistic and aesthetically pleasing productions, and users are more interested in personalized products. Thus, this study plans to use computer-aided means of artificial intelligence to investigate if it can help users create and generate more quality therapeutic music based on music that has therapeutic effects.

The object of the project is music, so there are also many challenges. First, music requires a temporal model. Second, music is usually composed of multiple instruments/tracks. They have their own temporal dynamics, but in general, they develop interdependently over time. Finally, in polyphonic music, notes are usually divided into chords or melodies. Therefore, the introduction of notes in chronological order is not necessarily harmonious or melodic, which is the problem to be addressed in this project.

Therefore, three dimensions need to be discussed in this study: whether the generated content is melody or accompaniment; whether single-track or multi-track music is generated; and whether the network architecture is CNN (convolutional neural network) or LSTM (recurrent neural network), or GAN (generative adversarial neural network) with CNN as the underlying architecture, which performs well in various image video tasks.

II. Prior Research on Composition by Artificial Intelligence

One simple technique in algorithmic composition is to select notes according to the order of the transition table. Generally it uses pattern recognition generation techniques, such as Markov Model and neural networks.

Back in 1994, Mozer proposed a recurrent self-predictive connectivity network for composing melodies with harmonic accompaniment. This network is called CONCERT. The network is trained on a set of fragments, with the aim of extracting style rules. CONCERT can then be used to compose a new piece of music.

Lichtenwalter concluded that most of the music generation methods for pattern recognition usually have mediocre performance and limited generality. Therefore, they propose to use sequential machine learning techniques with sliding windows to generate classifiers corresponding to the training set of music data. This approach has the advantage of greater generality than explicitly specified music syntax rules and has the potential to apply a variety of powerful and available non-sequential learning algorithms.

In 2016, Choik's team introduced a new approach to automating composition with text-based Long-Short Time Memory (LSTM) networks. In two case studies, the networks were designed to learn relationships in text documents of notes that represent chord progressions and drum tracks. In the experiments, a text-based recurrent neural network showed excellent learning effects for both cases, while the character-based network learned only chords, helping people compose by controlling the diversity parameters of the model.

The above-mentioned studies have given us a lot of inspiration and room for thought in our research. For example, music generation is a temporal model, and notes are usually classified as chords, harmonies or melodies. Therefore, it is not harmonious to introduce notes in chronological order, and a non-sequential learning algorithm needs to be designed to improve the generalization performance of music generation.

III. Introduction to the Algorithm

A. LSTM

Recurrent neural networks are specialized in processing temporal information, no matter such information is textual or pure digital data. The three mainstays of the recurrent neural network family are: recurrent neural network (RNN), long-short time memory model (LSTM), and gated recurrent unit (GRU). As is shown in Fig 1.

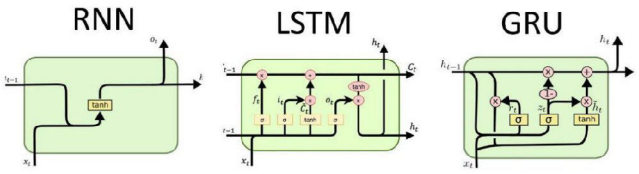


Fig. 1. Schematic diagram of recurrent neural networks

The standard LSTM model is a special type of RNN, with four special structures in each repetitive module, interacting in a special way. As in Figure 2, xt is the input of the current moment, ct-1 is the memory of the previous moment, and ht-1 is the output of the previous moment. LSTM is implemented in three gates: input gate, forget gate, and output gate.

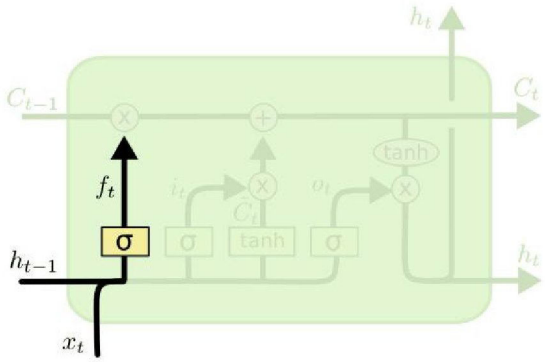


Fig. 2. Diagram of LSTM forget gate

The forget gate is used to calculate how much of the memory of the previous moment is preserved. The calculation equation is shown in Equation 1, which is calculated as a number between 0 and 1. As is shown in Fig 3.

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

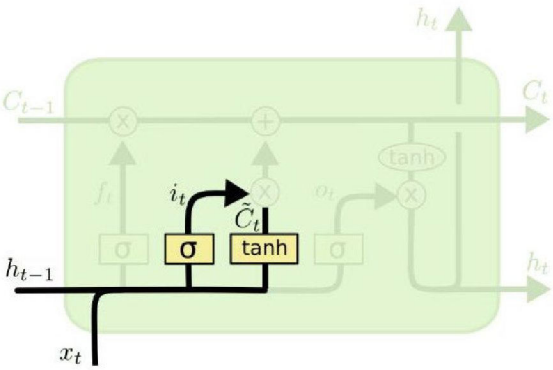


Fig. 3. Diagram of LSTM input gate

The input gate is used to calculate the memory of the current state, which is calculated as follows:

$$\begin{aligned} i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\ \tilde{C}_t &= \tanh\left(W_C \cdot [h_{t-1}, x_t] + b_C\right) \end{aligned} \tag{2}$$

The updated status at this time is Cnow = ~Ct*it, Ct = Ct-1*ft+Cnow.. As is shown in Fig.4.

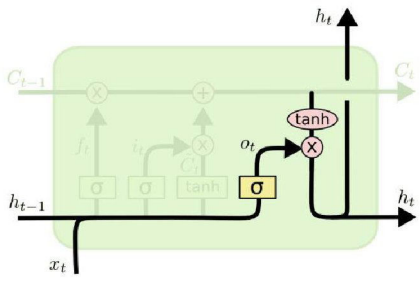


Fig. 4. Diagram of LSTM output gate

The output gate controls the output at this moment and is calculated as follows:

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o)$$
$$h_t = o_t * \tanh(C_t) \quad (3)$$

It is worth noting that in LSTM, both sigmoid and tanh activation functions are used. The tanh function has a larger gradient when the input is near 0 compared to the Sigmoid function, which usually leads to faster convergence of the model.

B. GAN

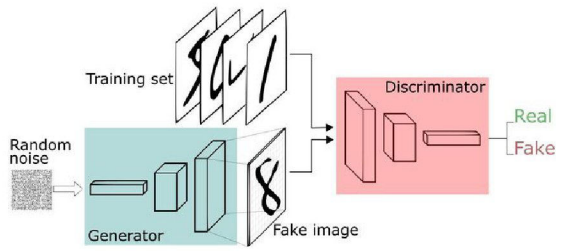


Fig. 5. GAN diagram

As shown in Figure 5, the simplest GAN (Adversarial Generative Network) consists of two parts, G and D. The generator tries to generate data from some probability distribution. The discriminator is like a jury that determines whether the input is from the generator or the real training set.

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))] \quad (4)$$

The adversarial modeling framework can be applied most directly when the models are all multilayer perceptrons. To learn the generator distribution P_g over the data x, I first defined a priori input noise variable P_z(z), and then mapped it into the data space according to G(z;θ_g), where G is the differentiable function characterized by the multilayer perceptron. We also needed to define a second multilayer perceptron D(s;θ_d), whose output is a single scalar. (x) denotes the probability that x is derived from real data and not from P_g. We trained D to maximize the probability of correctly assigning real and generated samples, so we can train G simultaneously by minimizing log(1-D(G(z))). That is, the discriminator D and the generator play a minimax game on the value function V(G,D).

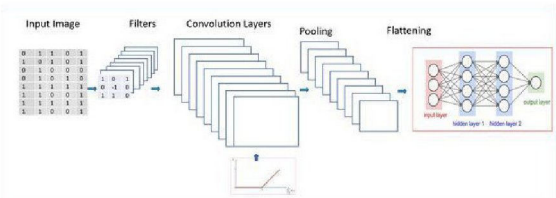


Fig. 6. CNN diagram

Currently, most of the underlying perceptrons of GANs use the CNN framework, i.e., convolutional shared computation method. The convolutional layer of the CNN is mainly convolutional operation, the input data is a matrix, and the features are extracted by multiple shared weight kernels/filter, and nonlinear steganography is performed. The structure of the CNN is shown in the larger figure in Figure 6, and the normal fully connected neural network is also shown as a control in the smaller figure on the right side of the figure.

C. Introduction to Therapeutic Music Genres

Genres of music in music therapy: Generally speaking, there are three genres of music in terms of the cause of treatment: calming and relaxing music, depression-relieving music, and brain-awakening and educational music.

Calming and relaxing music has a soft and lyrical melody that calms people down and relaxes the body and mind. It is suitable for people who suffer from insomnia, mental stress or cardiovascular disease. Calming and relaxing music can be divided into two series: calming and sleeping and stress-relieving and relaxing.

Depression-relieving music is divided into two categories, allowing patients to listen to sad music first and then to cheerful music. The therapeutic effect of depression-relieving music is to divert or vent anxiety with sad music, to cheer up the spirit, to listen to cheerful music to cause pleasant associations or good memories, to overcome depression or anxiety. Depression-relieving music is suitable for patients with depression, anxiety and neurasthenia.

The reason why brain-awakening and educational music can be educational is that it has different effects depending on the target. In terms of therapeutic effects, the stimulation of excitatory music is effective for people with low intelligence and low cortical arousal. The stimulation of attentive music can improve concentration and general sensory ability, and increase learning and work efficiency. Inspirational music can refresh the mind, eliminate fatigue, and stimulate inspiration.

D. Music Interpretation

Music has a hierarchical structure, with higher-level building blocks consisting of smaller recursive patterns. A song can be divided into multiple paragraphs, each paragraph has multiple phrases, each phrase has 4 bars, each bar has 4 beats, and each beat can be divided into 24 pixels. People pay attention to structural patterns associated with coherence, rhythm, tension, and emotional flow when listening to music. Therefore, it is crucial to explain the mechanisms of the temporal structure. As is shown in Fig.7.

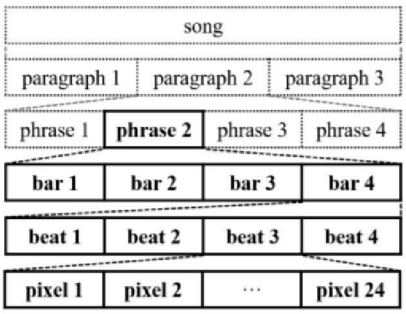


Fig. 7. Music composition

E. Introduction to the Data Set

The data set is divided into two parts, one is the public piano dataset, and the other is the multiple self-selected piano pieces.

The Lakh Pianoroll Dataset (LPD) is a collection of 174,154 multitrack piano rolls derived from the Lakh MIDI Dataset (LMD).

IV. Experiment and Analysis

In this section, the data needed to be preprocessed into input patterns acceptable to the computer network model, and then two deep neural networks were designed to compare the harmony of the two models in generating music.

A. Data Preprocessing

To convert a midi file into a "language", a music "fragment" should be defined, where a music fragment is equivalent to a word in a language. The music in the data set was cut into segments of the same length, with no overlap between them, and each segment was one beat long. The length of each beat was estimated by the MIDI Toolbox, and the length of each beat can vary from segment to segment. All high pitch levels in the segments were retained, where high pitch levels refer to pitches that do not contain scale information.

The diagram below shows the first bar of Chopin's Op. 67, No. 4, Mazzuca No. 47 in A minor, and shows how the length of the fragment is determined. Here, the length of a beat is one quarter note. As is shown in Fig .8.

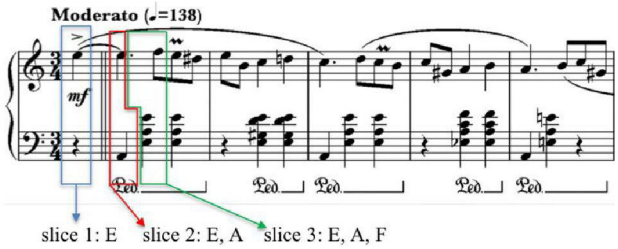


Fig. 8. Diagram of music data analysis

For the public data, we set the height to 84 to cover from C1 to C8. Thus, the size of the data tensor for each bar is 96 (time step) × 84 (note) × 5 (track). We consider four bars as one phrase, and trim the longer bars to the appropriate size accordingly. The final dataset has a total of 127,731 bars, and the goal of the model is to generate a five-track piano roll of four bars.

B. LSTM Results and Discussions

The LSTM model was built with 4 layers and the number of neurons were 30, 30, 64 and 1. To enhance the generalization performance of the model, a dropout layer was set after each of the first three layers of the LSTM. The dropout can randomly leave a portion of neurons in each round without working, where the random value is set to 50%. ReLU(y=max(0, x)) was used for each layer of neurons to provide a nonlinear mapping.

The network was trained 20, 40, 60, 80, 100, and 200 times on the dataset, and the reference values and the results of 100 and 200 times were sliced randomly in two segments for comparison. It is evident that after 200 times of learning, the LSTM model is better than 100 times of learning for the music pattern.

C. GAN Results and Discussions

GAN itself is a game-like network. With the above LSTM results, we upgraded its implementation effects and tested whether multi-track music can be generated. The schematic diagram of the built model is as follows. The music length is pre-defined, and the input is a set of random noise to generate a segment of a bar length, which is generated by the discriminator after the judgment.

The GAN superparameters are as follows: the maximum number of iterations was set to 50,000, the batch size was set to 64, the initial learning rate was set to 0.001, and the Adam optimizer (beta1=0.5, beta2=0.9) was chosen.

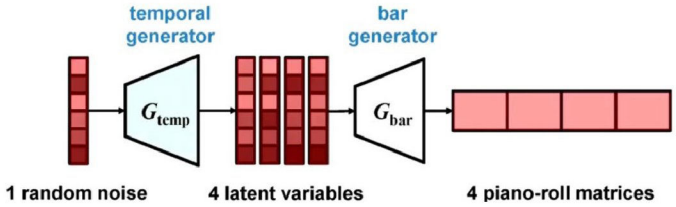
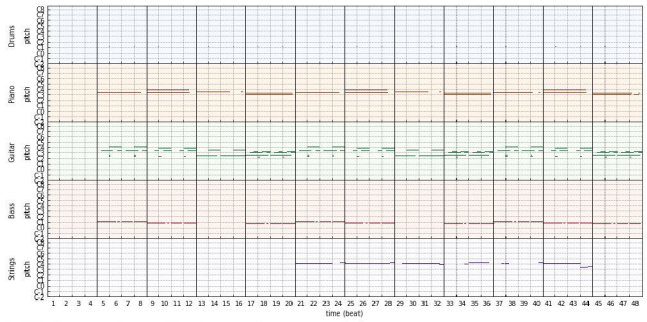
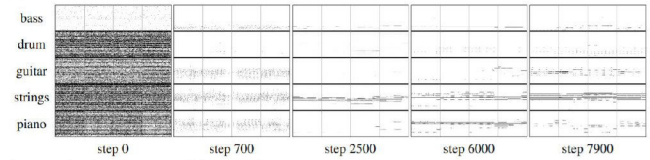


Fig. 9. Diagram of GAN music generator

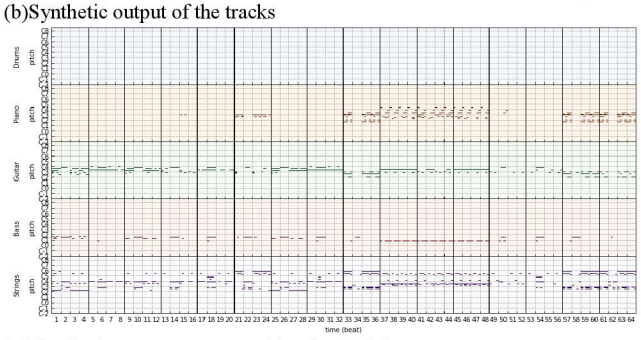
As is shown in Fig.9. The data distribution and generated results for each real track are shown in Figure 10(a) and (b). Except for the Drums, the chords of the other instruments are well articulated, with few interruptions, and the drum beats are generally distributed in a rhythmic pattern.



(a)Reference input for the tracks



(b)Synthetic output of the tracks



(c)The final sequence generated by the model

Fig. 10. GAN multi-track generation results

Figure 10(a) shows the output of a random sequence in the dataset, with the horizontal and vertical coordinates indicating time and notes, respectively. It can be seen that the different tracks of the midi sequence of the piece follow a specific pattern, especially for the tracks that need to generate melodies such as guitar and piano. Therefore, the final desired output of the model should be similar to the

sequence diagram. Analyzing Figure 10(b), we can see that at the beginning of the training, the audio midi sequence generated by the generator is unordered noise. After 700 epochs, the generated sequence is relatively orderly but still has a lot of noise, while after 7900 epochs, the generated sequence has a little noise but basically meets the expectation. After 30,000 epochs, the model can already produce very rhythmic music with very little noise, as shown in Figure 10(c).

V. Conclusion

This project investigated the scope and object of music therapy with therapeutic music as the object. The possibility of using computer-aided technology to replace manual music generation was analyzed. The experimental platform was built on two deep neural network models: LSTM and GAN. The experimental results showed that the network can generate many pieces of novel music at low cost and fast after sufficient learning of the information and patterns of music anticipation. It paves the way for applying artificial intelligence music therapists to people's lives, professional counseling organizations, and even hospitals.

References

[1] Bruscia, K. E. . Defining music therapy. Spring House Books, 1998.

[2] F Baker, and T. Wigram . Songwriting: Methods, Techniques and Clinical Applications for Music Therapy Clinicians, Educators, and Students. Jessica Kingsley, 2005.

[3] Hooper, J. . "Receptive Methods in Music Therapy: Techniques and Clinical Applications for Music Therapy Clinicians, Educators and Students." Music Therapy Perspectives 25.2(2007):127-129.

[4] Rolvsjord, R. . "Sophie Learns to Play her Songs of Tears– A case study exploring the dialectics between didactic and psychotherapeutic music therapy practices." Nordic Journal of Music Therapy (2001).

[5] Yuji K , Makoto S , Akihisa O , et al. Distinction of Students and Expert Therapists Based on Therapeutic Motions on a Robotic Device Using Support Vector Machine[J]. Journal of Medical and Biological Engineering, 2020:1-8.

[6] Wigram, T. , N. Pedersen , and L. Bonde . "A Comprehensive Guide to Music Therapy: Theory, Clinical Practice, Research and Training." (2002).

[7] Agold, C. M. , et al. "Music therapy for depression." Cochrane Database of Systematic Reviews 4(2003).