Investigating MIDI data simplification by AI models

Braden Ou

The Quarry Lane School, Dublin, CA 94568, the U.S.

*bradeno@andrew.cmu.edu

Abstract. According to the Smithsonian Institution, the art of making music has existed for over 35,000 years. As musical technology has improved, the music of the time has also improved and adapted to the new technology. In the recent expansion of technology from generative AI, text and image generation have become not only possible but also competitive with human-created text and images. As such, the development of AI-generated music is increasingly sparking considerable interest among musicians and developers alike, raising questions about the potential of AI to enhance or even replace human musical creativity. This paper will first explore the advancements of AI-generated music. Next, it will delve into the technologies and methodologies involved in generating music, as well as its current limitations using a basic LSTM (Long Short-Term Memory) model. Finally, it will explore the implications of this music for the whole music industry. By examining these various facets of AI-generated music, this research provides insights into AI's potential role in shaping the future of music. According to the analysis, a rudimentary AI model trained on complex music can produce music that is fairly elementary. Overall, these results shed light on guiding further exploration of the interaction between artificial intelligence and music.

Keywords: artificial intelligence, machine learning, music.

1. Introduction

The intersection of music and computing is a fascinating domain, with its origins dating back to the mid-20th century. This multifaceted field emerged when Australian computer scientist Trevor Pearcey used CSIRAC (Council for Scientific and Industrial Research Automatic Computer), Australia's first digital computer, to experiment with music in 1951. This seminal endeavour laid the groundwork for the digital production of sound, marking the birth of computer music [1, 2]. A significant breakthrough in computer music occurred in the late 1950s by Max Mathews at Bell Labs in the United States. He developed the MUSIC series, a collection of programming languages designed to generate music, which ushered in a new era in this field. MUSIC I, the first software of its kind, produced the inaugural musical note on a computer, catalysing the development of more sophisticated versions, notably MUSIC IV." The 1960s saw the emergence of computer music research centres in academic institutions. Prominent examples include Stanford's CCRMA (Centre for Computer Research in Music and Acoustics) and Princeton's Composers Inside Electronics. The following decades witnessed the commercialization of computer music with companies such as Moog and Fairlight introducing synthesizers. The 1980s also brought forth the Musical Instrument Digital Interface (MIDI), a significant advancement in music technology, which enabled communication between computers and synthesizers. The 21st century introduced a new dimension to computer music with the advent of machine learning and artificial intelligence. Algorithms capable of autonomously generating music were developed, exemplified by David Cope's "Experiments in Musical Intelligence" and OpenAI's "MuseNet" [1-3].

From its humble beginnings as an experimental offshoot of computer science and music, computer music has evolved to become a fundamental aspect of contemporary music production and performance. It has transcended its academic origins to permeate popular culture and continues to redefine the boundaries of musical creativity. Despite these advances, the field remains an open canvas for exploration and innovation, with researchers and artists continually pushing the envelope in terms of its potential.

Current music generation models use a specific model called Long Short-Term Memory (LSTM) [4, 5]. LSTMs, a type of Recurrent Neural Network (RNN) designed to learn from a long range of dependencies in sequential data, have shown promise in various applications, including natural language processing [6]. Several studies have showcased the potential of LSTM networks in generating coherent and aesthetically pleasing compositions by capturing complex structures and temporal dependencies in music. Huang et al. have demonstrated the effectiveness of generative adversarial networks (GANs) in generating multi-track music, provided a comprehensive review of deep learning techniques for music generation, including LSTMs, GANs, and Variational Autoencoders (VAEs) [7], and discussed the fundamentals of LSTM networks, highlighting their ability to learn long-range dependencies in sequential data [8]. Some studies have also applied LSTM networks to generate melodies and piano music, respectively, showcasing the potential of LSTMs in music generation, and exploring the multi-modal and hierarchical approaches to improve generated music, with successful applications in generating polyphonic and harmonically rich music [9]. Evaluating the quality of generated music remains a challenge, as highlighted by Nakamura et al., who emphasized the importance of combining quantitative metrics with qualitative listener feedback [10]. Lastly, the ethical and practical implications of AI-generated music were discussed by Frontiers in Artificial Intelligence, highlighting the need to consider the impact of AI on human creativity, authorship, and the future of the music industry [11].

In order to showcase the effectiveness and limitations of artificial intelligence and music, this paper will build upon preexisting models and modify them. To be specific, this study will utilize an LSTM model that generates music, and modifying it based on specific sets of data to create an experiment to test similarity. Hence, one will look over existing research, as well as building the own LSTM model to test the limitations ourselves. The rest part of the paper is organized as follows. Section 2 will be a description of the data. Section 3 will be the results and a discussion of the results. Section 4 will be about the limitations and prospects of the research. Finally, Section 5 will be a conclusion that details the potential implications of this research.

2. Data & method

The LSTM model is an exceptional tool for music generation. LSTM models, with their aptitude for handling long-range dependencies, can capture patterns from a broad range of time steps. This feature is particularly advantageous for music generation, where motifs, themes, and harmonic progressions frequently recur throughout a composition. Therefore, a system that can accurately model these patterns is crucial. A graphical representation of the LSTM model is provided in Figure 1. Here, xt is the input, ht-1 is the output of the previous layer, and ct-1 is the previous cell state. The distinguishing feature of LSTM models is their gating mechanisms which regulate the flow of information, allowing the model to selectively remember and forget information based on its relevance. Specifically, there are three types of gates embedded in LSTM models: Input gate, Forget gate, and Output gate. The Input gate decides which new information will be updated into the cell state. This gate is useful for introducing new elements into the music, such as changes in pitch, duration, harmony, or other relevant attributes. The Forget gate, on the other hand, decides which parts of the previous cell state should be discarded. The Forget gate uses a sigmoid function to look for values to forget. It could be used to forget patterns that were only prevalent for a short period of time or no longer serve the current musical context. Lastly, the Output gate decides what information from the cell state should be conveyed as the output of the LSTM cell. This gate is fundamental in determining which information from the cell state should be used to generate the output at the current time step, thus influencing the present moment in the music composition. In Figure 1, the Input gate corresponds to the leftmost sigmoid layer. It takes in the value from the previous output ht-1 and a new set of inputs xt, and using these two values, it outputs a number between 0 and 1 for the values in the previous cell state of Ct-1. The greater the number that is outputted, the more likely the model will keep that value. Let's assume that the values to be removed will be stored in a forget gate, denoted as Ft. The purpose of the middle two arrows, the sigmoid and tanh(x) arrows, is to find what new information should be inputted into the cell state. The sigmoid function is first used to determine what values should be updated in the cell state, and then the tanh(x) layer will find possible values that could be added to the state. Then, the model combines these two layers to update the cell state. Let's denote this combination as Ut. Now that the decision of how to update the cell states with Ft and Ut is made, the model can start updating the cell state. First, it will multiply the old state Ct-1 by Ft, and then it will add Ut to the cell state to create the new cell state. Finally, one needs to determine what the output ht will be in this cell state. First, the cell state will be put into a sigmoid layer to determine which parts of the cell state will be output, and it will also be separately put into a tanh(x) function, Then, the model will multiply the outputs of both cell state transformations to determine which parts the model will output.

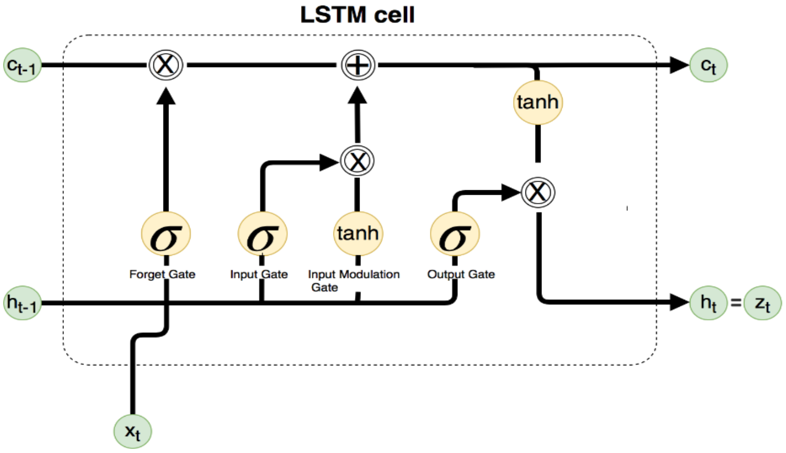

Figure 1. A sketch of LSTM cell [12].

This study will primarily employ an LSTM-based music generation model. In the application of the LSTM model, this study plans on utilizing pitch, duration, and step as the primary parameters that the model will interpret and learn from. With these features, it is hoped that the LSTM model can generate music that is not only rich in detail but also respects the fundamental elements of musical composition. The LSTM model will be trained on the Maestro Dataset, a large collection of MIDI and audio files. Collected from the International Piano-e-Competition and a part of the Magenta research project, these files are complex piano pieces played by virtuoso pianists.

3. Results & discussion

The results from the experiment offer intriguing insights into the capability of LSTM models for music generation and contribute to the understanding of the challenges and prospects in this research domain. As shown in Figure 2, observations from the normalized training loss graph indicate that the model learned effectively over the training epochs. The loss value gradually decreased and stabilized around the range of 0.1, signifying that the model was able to generalize well on the training data and minimize the difference between the actual and predicted sequences. Notably, an abnormal spike was observed at the 105th epoch, causing a sudden increase in the loss value. However, this spike was temporary and the loss quickly decreased in subsequent epochs. This spike might be attributed to several factors, such as learning rate schedules, model instability at that point in training, or potential outliers in the batch of data used during that epoch.

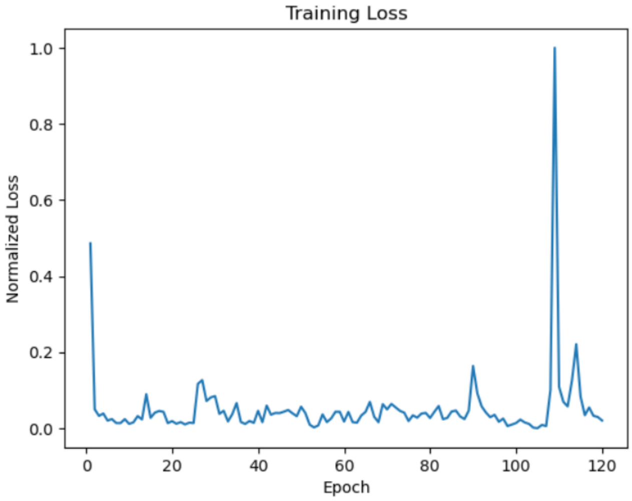

Figure 2. Loss as a function of Epoch (Photo/Picture credit: Original).

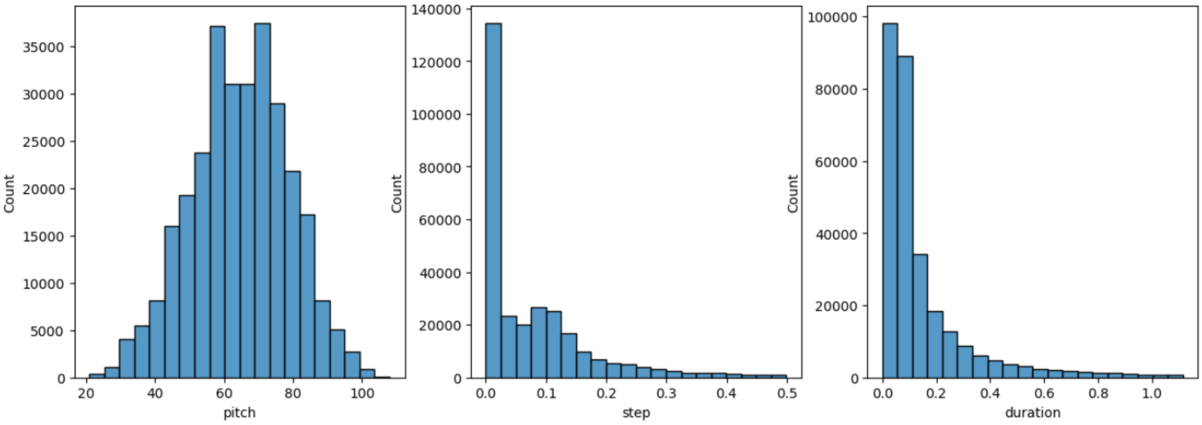

Figure 3. Distribution of training (Photo/Picture credit: Original).

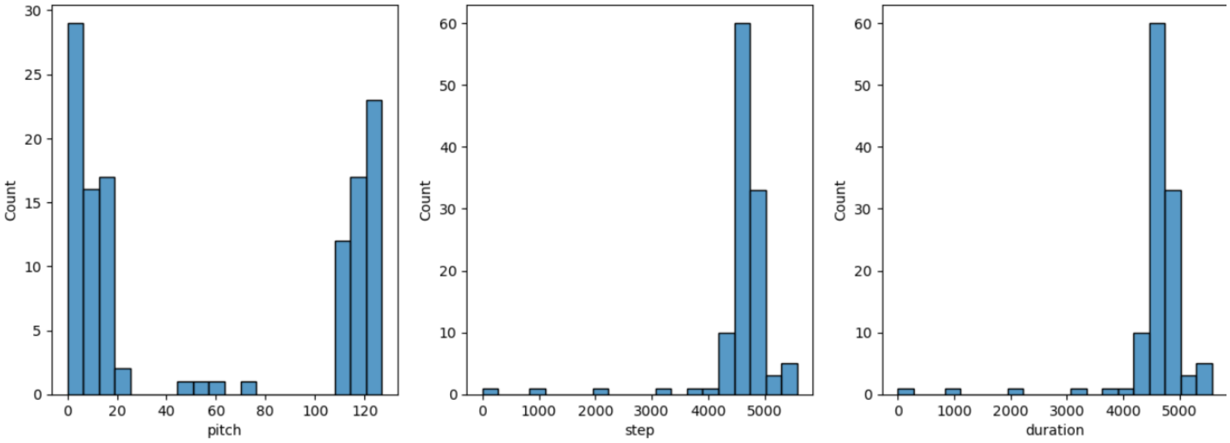

Figure 4. Distribution of testing (Photo/Picture credit: Original).

Seen from Figure 3, an analysis of the training data from the Maestro dataset reveals a fairly normalized distribution of pitches. The step and duration characteristics, however, demonstrated a right skew. This indicates that shorter notes and rests are more prevalent in the training data, which is typical for many genres of music, particularly those that rely heavily on rhythmic intricacy. In contrast, as depicted in Figure 4, the pitch distribution of the generated sequences exhibited a bimodal pattern, clustering at the extreme ends of the pitch ranges. Additionally, the steps and duration characteristics were left-skewed. The bimodal distribution of the pitch could be attributed to the model's tendency to simplify complex polyphonic input into the two ends of the spectrum of possible MIDI notes. This interpretation aligns with the constraints of the LSTM model used, which was designed to handle only monophonic notes. This simplification by the LSTM model, while potentially a limitation in capturing the full richness of the input music, provides an interesting insight into how artificial intelligence can manage complexity. Despite its limitation in handling polyphonic music, the model was still able to generate musically coherent outputs. This observation opens up exciting avenues for future research, especially in understanding and improving how LSTM and similar models handle and generate polyphonic music.

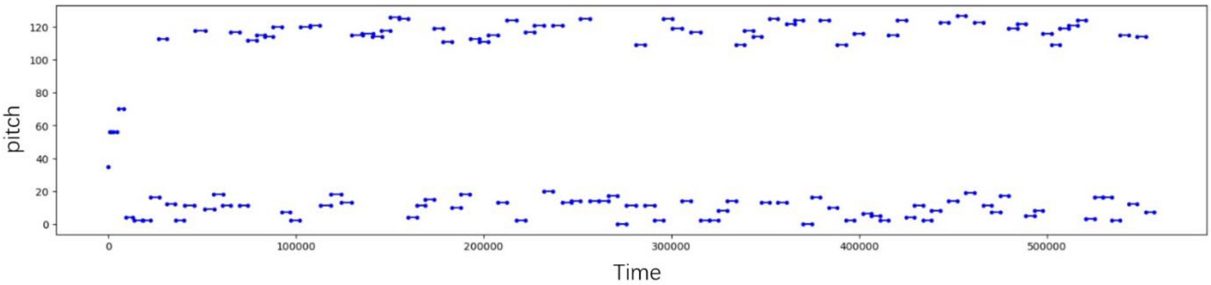

Figure 5. Output results (Photo/Picture credit: Original).

Despite the odd distributions of pitches, steps, and duration in the generated notes, as illustrated in Figure 5, the piano roll shows that the track itself has varied pitches, steps, and duration, showing how despite the bimodal nature of the output, the output itself is still varied and unique. In summary, the LSTM model demonstrated promising results in generating music, successfully learning from the training data, and synthesizing new musical sequences. However, the model's tendency to simplify complex inputs, evident in the pitch distributions of the generated sequences, poses a challenge that future research in this field should aim to overcome.

4. Limitations & prospects

The present research has sought to leverage the power of LSTM models for the generation of music. While the study has achieved several significant findings, it has also faced a set of limitations that pave the way for future research opportunities. One notable constraint was the computational power available for the project. LSTM models, being recurrent neural networks, are computationally intensive due to their inherent sequential nature. Unfortunately, the lack of substantial GPU resources imposed a considerable limitation on the research. GPU acceleration is essential for training deep learning models, as it significantly reduces the time required for the training process. The lack of such computational power resulted in a compromise to the complexity of the music generation process. As a workaround, the research employed monotonous elements of MIDI songs to reduce the complexity and dimensionality of the input data. While this approach allowed the LSTM model to function within the available computational constraints, it inevitably reduced the variation and richness of the generated music. Essentially, the simplification of input might have led to an oversimplification of the model's understanding of music, reducing its capacity to generate music with complex patterns and structures.

Interestingly, the unique constraint that was intentionally introduced to the model involved training it with complex polyphonic music, while the LSTM model was geared to handle only monophonic notes.

This approach intended to observe how the AI could simplify the computational process, and although it might initially appear as a limitation, it was indeed a key research point. In effect, the model had to learn to distil the complexity of polyphonic music down to a single melodic line, which was an interesting method for assessing the model's capability to handle complex musical structures.

Despite these limitations, the study has promising prospects. With advancements in technology and more accessible computational resources, future research could overcome the constraints faced by the present study. Increased computational power would allow for the training of more complex LSTM models capable of understanding and generating music with higher variation and intricacy. Moreover, it would permit the utilization of more complex and nuanced MIDI song elements, capturing more intricate patterns and generating richer music.

The present research can be further expanded in several ways. Future studies could investigate the integration of other music-related parameters, such as tempo, dynamics, and articulation, which would provide a more comprehensive musical understanding of the model. In addition, future research could explore the fusion of different architectures or the use of more recent models, such as Transformer or Capsule Networks, to see whether they can improve performance and generate more musically coherent output. Lastly, the use of ensemble methods and the introduction of various forms of regularization could also be looked into for better model performance. Overall, while the limitations of the current project are clear, they open a wide range of exciting opportunities for future research in music generation using deep learning.

5. Conclusion

This study delved into the fascinating realm of music generation using deep learning, specifically leveraging the power of Long Short-Term Memory (LSTM) models. The experiment demonstrated the ability of LSTM models to learn from complex sequential data, as evidenced by the training loss graph, and successfully generate musically coherent sequences. However, the generated output, especially the pitch distributions, underscored the model's inherent limitations, particularly its propensity to simplify complex polyphonic input into more manageable monophonic output.

Despite these limitations, the findings provide invaluable insights into how artificial intelligence navigates complexity. The process of simplification carried out by the LSTM model, while limiting in some respects, illustrates a unique mechanism through which AI can manage intricate structures. This discovery opens the door to myriad research opportunities, particularly aimed at enhancing how LSTM and similar models process and generate polyphonic music. The observation that AI can simplify complex polyphonic music into more manageable patterns has significant applications for music composers and producers. They could utilize AI algorithms as a tool to generate foundational structures, melodies, or chord progressions, thus expediting the composition process and allowing more time for creative exploration in other musical dimensions. Such a development could democratize music production, empowering individuals with limited technical training to craft complex musical pieces. Moreover, the implications of AI in music generation extend to music pedagogy. As AI models advance and become more accessible, they could be incorporated into music curricula, providing learners with experiential knowledge at the intersection of technology and music, and preparing future musicians for the evolving demands of the music industry.

Nevertheless, the advent of AI in music brings to the forefront ethical and copyright considerations. As AI algorithms begin to produce music akin to human compositions, questions of authorship and royalty rights are likely to become increasingly pertinent. This necessitates a rigorous dialogue among policymakers, legal experts, and industry stakeholders to frame appropriate regulatory measures. While some may perceive the emergence of AI-generated music as a threat to traditional musicianship, it should be seen as an opportunity for musicians to harness new creative frontiers. Rather than supplanting human creativity, AI can serve as a complementary tool to stimulate musical imagination, enabling artists to traverse novel musical landscapes.

To sum up, while there are hurdles to overcome, the realm of music generation using deep learning exhibits immense potential. This study represents a small step in an exciting journey toward the harmonious merger of technology and art, where creativity can be expressed and explored in entirely new dimensions. The quest to refine the use of AI in music will undeniably continue to be a fascinating and transformative journey, offering novel ways to understand and appreciate the universal language of music.

References

[1] Fulzele P, Singh R, Kaushik N, et al. 2018 A hybrid model for music genre classification using LSTM and SVM. 2018 Eleventh International Conference on Contemporary Computing (IC3). IEEE, pp 1-3.

[2] Art & Music. The Smithsonian Institution's Human Origins Program, 19 Sept. 2022, Retrieved from: https://humanorigins.si.edu/evidence/behavior/art-music#:~:text=Making%20music%20is%20a%20universal,at%20least%2035%2C000%20years%20ago.

[3] Avdeeff M 2019 Artificial intelligence & popular music: SKYGGE, flow machines, and the audio uncanny valley. Arts. MDPI, vol 8(4) p 130.

[4] Computer Music (so Far). Short History of Computer Music, Retrieved from: https://artsites.ucsc.edu/EMS/Music/equipment/computers/history/history.html

[5] Kotecha N and Young P 2018 Generating music using an LSTM network. arXiv preprint arXiv:1804.07300.

[6] Mangal S, Modak R and Joshi P 2019 Lstm based music generation system. arXiv preprint arXiv:1908.01080.

[7] Moffat D and Sandler M B 2019 Approaches in intelligent music production. Arts. MDPI, vol 8(4) p 125.

[8] Saxena S 2023 Learn about Long Short-Term Memory (LSTM) Algorithms. Analytics Vidhya, 3 Mar. Retrieved from: https://www.analyticsvidhya.com/blog/2021/03/introduction-to-long-short-term-memory-lstm/.

[9] Carnovalini F and Rodà A 2020 Computational creativity and music generation systems: An introduction to the state of the art. Frontiers in Artificial Intelligence, vol 3 p 14.

[10] Sturm B L T, Iglesias M, Ben-Tal O, et al 2019 Artificial intelligence and music: open questions of copyright law and engineering praxis. Arts. MDPI vol 8(3) p 115.

[11] Ycart A and Benetos E 2017 A study on LSTM networks for polyphonic music sequence modelling. ISMIR, p 11.

[12] Zhang R, Liu Y, and Sun H 2020 Physics-informed multi-LSTM networks for metamodeling of nonlinear structures Computer Methods in Applied Mechanics and Engineering, vol 369 p 113226.