Comparative Study of LSTM and Transformer Models in Music Generation

Kasper Thomas Gartside Knudsen

April 1, 2025

Contents

1 Abstract

This research undertakes a comparative study of LSTM and Transformer models in the realm of AI-generated music, with an emphasis on balancing musical novelty and coherence. The study extends to explore the integration of advanced audio generation models like WaveNet and SampleRNN, aiming to enhance the quality of generated music. We investigate the capabilities of LSTMs compared to state-of-the-art Transformer models, examining their performance across different datasets and their potential synergy with advanced audio synthesis techniques.

2 Motivation

Drawing from my Bachelor's project titled "Domain Coverage of Low-Resource Chatbots," where I extensively investigated the capabilities of LSTMs in chatbot development, I'm keen to transition this exploration to the realm of music generation. While much work has been done on individual architectures, like LSTMs or Transformers, a comparison that takes into account multiple datasets, architectures, and how they work together is less explored.

LSTMs have historically been effective for sequence-based tasks due to their memory capabilities. Choi et al.'s tutorial on deep learning for music information retrieval emphasizes this strength and lays a foundation for LSTMs' application in music. This introduces an open problem: Can advancements in LSTM tuning and dataset usage contest newer models in terms of musical coherence and novelty?

On the other hand, Transformers have recently gained significant attention, especially with successes like MuseNet highlighting their potential in music generation. Huang et al.'s research on the Music Transformer further emphasizes the capabilities of Transformers in this domain. This leads to another open problem: When considering efficiency, training, and output quality, how do LSTMs compare to Transformers?

In addition to exploring LSTM and Transformer models, this study also delves into the potential of integrating these models with advanced neural network architectures like WaveNet and SampleRNN. These models represent the forefront of audio generation technology and offer a pathway to significantly enhance the fidelity and realism of AI-generated music.

3 Research Objectives and Hypotheses

This study aims to achieve the following objectives:

1. To conduct a comparative analysis of LSTM and Transformer models in generating music, assessing their efficiency, musicality, and novelty.
2. To explore the integration of advanced audio generation models, WaveNet and SampleRNN, with LSTM and Transformer architectures, and assess their combined effectiveness in producing high-quality, genre-specific music.

Based on these objectives, the study hypothesizes that:

- While both LSTM and Transformer models are capable of generating coherent music, their performance might vary significantly depending on the nature of the dataset (MIDI vs. audio).
- The integration of LSTM or Transformer models with WaveNet and SampleRNN could lead to enhanced music generation capabilities, surpassing the limitations of using these models in isolation.

4 Methodology

The project will conduct a comparative study between LSTM and Transformer models in music generation, examining their individual capabilities, strengths, and weaknesses in the context of music generation. Their performance and output quality will be evaluated with the use of quantitative evaluation metrics (such as Inception Score and Perplexity), investigating the influence of different data representations, including MIDI datasets (such as MAESTRO) and audio files. This will delve into how the structured nature of MIDI or the richer nuances of audio files may impact the generated music.

4.1 Genre-Specific Consideration

In the domain of music generation, genre remains a crucial yet complex dimension. While the core objective centers around comparing LSTM and Transformer architectures, a deliberate effort is made to segment and evaluate the datasets based on distinct genres. Datasets often encompass a mixture of genres, each with its distinct characteristics. To this end, models will be developed with the capability to discern between and generate music across these varied genres. This dimension introduces added complexity, from data segmentation challenges to the evaluation of genre authenticity.

4.2 Integration with Advanced Audio Generation Models

In exploring the capabilities of LSTM and Transformer models, we also consider their potential integration with WaveNet and SampleRNN. This involves examining how the structural understanding captured by LSTMs and Transformers can be combined with the high-fidelity audio generation capabilities of WaveNet and SampleRNN. The methodology will extend to evaluate the feasibility and effectiveness of this integration in producing high-quality, genre-specific music.

4.3 Data Handling and Analysis

For the LSTM and Transformer models, we will process the MIDI files from the MAESTRO dataset and audio files from the GTZAN dataset, extracting relevant features for music generation. The performance of these models will be evaluated based on metrics such as Inception Score and Perplexity.

For the advanced audio generation models, we will analyze their ability to synthesize realistic audio from the structural outputs provided by LSTM and Transformer models. The evaluation will focus on the audio quality, genre fidelity, and the seamless integration of these models with traditional architectures.

4.4 Challenges and Limitations

This study may encounter challenges, including:

- The computational complexity of training and integrating advanced models like WaveNet and SampleRNN.
- Potential difficulties in achieving seamless integration between sequence-based models (LSTM/Transformer) and raw audio generation models.
- Limitations in available data, particularly in representing all genres equally in the GTZAN dataset.

These challenges will be addressed through careful experimental design and resource allocation.

5 Proposed Experiments

- **Comparative Analysis:** Train both LSTM and Transformer models on identical datasets, analysing outputs for musicality, novelty, and coherence.
- **Dataset Influence:** Venturing into lesser-explored territory by training the models on MIDI files and audio, understanding dataset nuances' impact on music generation.
- **Model and Hyperparameter Fine-tuning:** To ensure optimal performance, experiment with different model configurations, hyperparameters, and training strategies.

6 Datasets

In this study, two primary datasets are utilized: the MAESTRO dataset and the GTZAN dataset. Each offers unique characteristics and challenges for the models being compared.

6.1 MAESTRO Dataset

The MAESTRO (MIDI and Audio Edited for Synchronous TRacks and Organization) dataset is a comprehensive collection of MIDI recordings. It includes over 200 hours of paired audio and MIDI recordings from ten years of International Piano-e-Competition events. The MIDI format allows for precise control over musical elements, making it a popular choice for training music generation models.

The structured nature of MIDI files in the MAESTRO dataset provides a clear framework for understanding music generation. These files contain explicit information about notes, timing, velocity, and other musical dynamics, making it easier for AI models to learn patterns and structures in music.

6.2 GTZAN Dataset

The GTZAN dataset, on the other hand, is a collection of 1000 audio tracks evenly distributed across 10 genres, each 30 seconds long. This dataset is widely used for tasks

like genre classification and music analysis. Unlike MIDI files, raw audio presents a richer, more nuanced view of music, encompassing timbre, tone, and the complexity of real-world sound recordings.

Switching to the GTZAN dataset represents a shift in approach from a structured, note-based analysis to a more holistic, audio-based perspective. This move could unveil new insights into how AI models handle the complexity and subtleties present in raw audio data, which is more representative of how humans experience music. The challenge, however, lies in the increased computational complexity and the need for models to interpret a far more varied and less structured form of data.

6.3 Comparative Analysis Using Different Datasets

The comparison between LSTM and Transformer models using these two datasets offers a multifaceted understanding of their capabilities. While the MAESTRO dataset allows for an analysis of how well these models learn and replicate structured musical compositions, the GTZAN dataset presents an opportunity to explore their performance in a more complex and less predictable environment.

This dual-dataset approach aims to evaluate not just the technical efficiency of each model in processing different types of data, but also their ability to generate music that is both novel and coherent, reflecting the diverse and intricate nature of real-world music. The results could provide significant insights into the current state and future potential of AI in music generation.

7 Model Architectures

This section delves into the architectures of the Transformer models, with a spotlight on MuseNet and the Music Transformer, as well as the LSTM model which will be utilized in this study.

7.1 Transformer Models

Transformer models have garnered attention due to their self-attention mechanism that allows for long-term dependency modeling, making them a suitable choice for music generation tasks.

7.1.1 Music Transformer

Music Transformer, developed by Huang et al., leverages the self-attention mechanism of Transformer models to handle long-range dependencies crucial for generating music with coherent structure over time. The key innovation lies in a modified relative attention mechanism that reduces the memory complexity from quadratic to linear in the sequence length, enabling the generation of longer musical compositions. This adjustment allows the Music Transformer to generate minute-long compositions (thousands of steps), significantly longer than previous models. It can also generate coherent continuations elaborating on a given motif and, in a sequence-to-sequence setup, create accompaniments conditioned on melodies. The model demonstrated its effectiveness on two datasets, JSB Chorales and Piano-e-Competition, achieving state-of-the-art results on the latter.

7.1.2 MuseNet

MuseNet, developed by OpenAI, is a deep neural network adept at generating four-minute musical compositions incorporating 10 different instruments and blending a variety of musical styles from classical to contemporary. Unlike some models, it wasn't explicitly programmed with predefined musical understanding, but instead, it autonomously discerned harmony, rhythm, and style patterns by predicting subsequent tokens in an extensive collection of MIDI files.

The underlying architecture of MuseNet is a large-scale transformer model, similar to GPT-2, facilitating the prediction of the next token in a sequence, whether it's audio or text. MuseNet's design comprises a 72-layer network with 24 attention heads, which is instrumental in capturing long-term structural nuances in music, a significant advancement in music generation AI.

A distinctive feature of MuseNet is its capability to blend different musical styles convincingly. For instance, it can take the initial notes of a classical piece and generate a composition in a pop style with a variety of instruments. The training data for MuseNet is diverse, sourced from various collections encompassing classical, jazz, pop, and other genres, which significantly contributes to its versatility in music generation.

Furthermore, MuseNet employs composer and instrumentation tokens to allow more control over the generated music, enabling the conditioning of the model to create compositions in a chosen style, demonstrating a sophisticated level of flexibility and adaptability in music generation.

In comparison to other music generation models, the long-term structure retention and the ability to blend different musical styles seamlessly make MuseNet a noteworthy model in the realm of AI-generated music.

7.2 LSTM Models

Long Short-Term Memory (LSTM) models have been a staple in sequence modeling tasks due to their capacity to retain information over long sequences, which is crucial for tasks like music generation. Their architecture allows for the storage and retrieval of information over long periods, making them a viable choice for generating coherent musical compositions.

7.3 Comparative Analysis of Model Architectures

The exploration of LSTM, Music Transformer, and MuseNet reveals distinct architectural approaches toward music generation. Each model presents a unique set of strengths and potential limitations.

- **Handling of Long-term Dependencies:** The Music Transformer employs a modified relative attention mechanism to address long-term dependencies, enabling the generation of longer compositions. In contrast, LSTMs use their memory cells to manage sequence information over time, while MuseNet's extensive transformer architecture captures long-term structural nuances in music.
- **Musical Coherence and Novelty:** LSTMs have shown promise in generating musically coherent pieces, albeit often with simpler structures. Music Transformer

and MuseNet, with their transformer architectures, have demonstrated the capability to generate more complex, novel musical compositions while retaining coherence over extended lengths.

- **Genre Versatility and Style Blending:** MuseNet stands out in its ability to blend different musical styles convincingly, a feature not explicitly demonstrated by Music Transformer or LSTMs.
- **Training and Computational Efficiency:** Training LSTMs might be less computationally intensive compared to the larger transformer models. However, the latter have shown to achieve remarkable results in music generation, often justifying the increased computational demand.
- **Control and Conditioning:** MuseNet employs composer and instrumentation tokens for controlled generation, a level of flexibility that is instrumental in generating a varied repertoire of music. This aspect of control and conditioning is an area that could be explored further in LSTMs and the Music Transformer.

The comparative analysis accentuates the evolving landscape of AI in music generation, showcasing how different architectures contribute to advancing the field. It also underscores the importance of continued exploration in model architectures and training methodologies to push the boundaries of what's achievable in AI-driven music generation.

8 Utilizing Advanced Audio Generation Models

In our pursuit to enhance the quality of AI-generated music, we explore the integration of advanced neural network architectures specifically designed for audio generation: WaveNet and SampleRNN. These models represent the cutting edge in synthesizing high-fidelity audio and have shown remarkable success in tasks such as speech synthesis and music generation.

8.1 WaveNet

Developed by DeepMind, WaveNet is a deep generative model of raw audio waveforms using a dilated convolutional neural network [**?**]. WaveNet's architecture allows it to model audio signals with high temporal resolution, capturing the intricate patterns necessary for producing realistic and coherent audio sequences.

8.2 SampleRNN

SampleRNN, developed by researchers at the University of Montreal, is a recurrent neural network-based model that processes audio samples at multiple temporal resolutions [**?**]. It is particularly noted for its ability to capture and generate complex audio textures, making it suitable for diverse music generation tasks.

8.3 Integration in Music Generation

Leveraging these advanced models presents an opportunity to significantly enhance the realism and quality of generated music. The proposed approach involves using the LSTM

or Transformer models to capture the high-level structure of music and then employing WaveNet or SampleRNN for the actual audio waveform generation. This hybrid method aims to combine the strengths of different architectures to produce rich and varied musical outputs.

9 Future Work

Future research could extend this study by:

- Exploring alternative architectures and hybrid models for music generation.
- Conducting user studies to evaluate the subjective quality of the generated music.
- Investigating the potential of these models in other applications, such as soundtrack generation for games or movies.

10 Discussion and Conclusion

The findings of this study highlight the strengths and weaknesses of LSTM and Transformer models in music generation. Furthermore, the exploration of integrating these models with advanced audio generation technologies like WaveNet and SampleRNN opens new avenues for creating highly realistic and varied musical pieces. This research not only contributes to the understanding of AI in music generation but also sets the stage for future explorations into hybrid approaches that combine the best of sequence modeling and audio synthesis technologies.

11 References

- Choi, K., Fazekas, G., Sandler, M., & Cho, K. (2016). A tutorial on deep learning for music information retrieval. arXiv preprint arXiv:1709.04396.
- Huang, C., Vaswani, A., Uszkoreit, J., Simon, I., Hawthorne, C., Shazeer, N., ... & Chen, D. (2018). Music Transformer. arXiv preprint arXiv:1809.04281.
- Payne, C. (2019). MuseNet. OpenAI Blog. Retrieved from https://openai.com/blog/musenet/
- Knudsen, K. T. G. (2022). Domain Coverage of Low-Resource Chatbots. Bachelor's Project.
- Hawthorne, C., Elsen, E., Song, J., Roberts, A., Raffel, C., Engel, J., ... & Eck, D. (2019). MAESTRO: A large-scale dataset and benchmarks for music transcription. In Proceedings of the International Society for Music Information Retrieval Conference. Retrieved from https://magenta.tensorflow.org/datasets/maestro.
- van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A., & Kavukcuoglu, K. (2016). WaveNet: A generative model for raw audio. arXiv preprint arXiv:1609.03499.
- Mehri, S., Kumar, K., Gulrajani, I., Kumar, R., Jain, S., Sotelo, J., Courville, A., & Bengio, Y. (2016). SampleRNN: An unconditional end-to-end neural audio generation model. arXiv preprint arXiv:1612.07837.