Deep Learning in Music Generation: A Comprehensive Investigation of Models, Challenges and Future Directions

Xiangchen Kong

Computer Science, University of California, Davis, 95616, Davis, CA, United State

Abstract. Deep learning has made a lot of progress in the field of music generation. It now has powerful tools for both preserving traditional music and creating new, innovative compositions. This review explores various and recent deep learning models, such as Long Short-Term Memory (LSTM) networks, Transformer-based models, Reinforcement Learning (RL), and Diffusion-based architectures, and how they are applied to music generation. LSTMs effectively capture temporal dependencies, which are vital for producing coherent melodies and chord progressions. Transformer models, like MUSICGEN and STEMGEN, handle large amounts of data and dependencies efficiently, but they need a lot of computational resources. Reinforcement Learning models, such as MusicRL, combine human feedback to fine-tune AI-generated compositions based on the individual's preferences. Diffusion-based models, like MusicLDM, enhance audio fidelity, though real-time application remains a challenge. The objective of emotion-conditioned models, such as ECMusicLM, is to combine music with emotional cues so that the output has a stronger emotional resonance. However, each model faces its own set of limitations, such as computational inefficiency, data dependency, and challenges in capturing complex emotional nuances. Future research should focus on improving the computational efficiency of these models, expanding training datasets, and integrating more interactive, real-time systems.

1 Introduction

Deep learning has revolutionized numerous aspects of creativity and one of the most interesting advancements is in the music generation. Artificial intelligence can now produce music that keeps traditional musical structures while also exploring new creative paths by using the complicated neural networks. Long Short-Term Memory (LSTM) networks have been shown to be exceptionally effective at modeling sequential data, like music, where the relationships between notes and chords in terms of time are very critical. For example, research on chord-based music generation using LSTM highlights how this architecture can be used to generate multi-style chord progressions while preserving the stylistic elements of diverse musical forms [1].

Corresponding author: xchkong@ucdavis.edu

Based on the LSTM, there are more advanced techniques, such as Bidirectional Long Short-Term Memory (Bi-LSTM), have been applied in music generation. A remarkable example is a study that focuses on the preservation and development of Xi'an drum music, a special form of Chinese folk music. The model uses Bi-LSTM combined with the Actor-Critic reinforcement learning algorithm to generate music while following the rules of music theory for harmony. This system makes it possible to write compositions that have the beauty of traditional music [2].

As deep learning develops, models like GPianoroll, which combines human feedback, make it possible for AI-made compositions to improve even better. GPianoroll uses Bayesian Optimization to fine-tune model parameters and ensure that the generated music could match user preferences. This makes it possible for AI and human composers to work together [3].

More recent developments and models such as Music ControlNet. They allow fine-tuned, time-specific control over musical features such as melody and rhythm. more control over musical elements than ever before, which lets them build more detailed and dynamic compositions [4]. Similarly, MusicLM, a model that is designed to generate music from text descriptions, has introduced an emotion-conditioning variant called Emotion-Conditioned MusicLM (ECMusicLM), which enhances emotional resonance in generated music by conditioning on both text and emotional cues [5].

Reinforcement learning has been very helpful in producing AI-generated music with human preferences. The MusicRL model incorporates Reinforcement Learning from Human Feedback (RLHF) to ensure that the generated music could reflect the subjective preferences of humans, particularly in areas like musicality and emotional tone [6]. This model, along with others like STEMGEN, which generates music based on a specific musical context, shows how AI can act as a collaborator in the creative process [7].

There have been more developments made to the efficiency and novelty of music generation. For instance, the Grey Wolf Optimizer has been applied to optimize LSTM hyperparameters, speeding up the training process and enhancing the quality of music generation [8]. Meanwhile, models like MusicLDM employ Stable Diffusion and AudioLDM architectures to generate novel compositions through beat-synchronous data augmentation techniques, which lets people make the creation of diverse, original music [9].

Finally, MUSICGEN, a single-stage Transformer Language Model, simplifies the music generation process by processing compressed discrete music tokens. It is an excellent alternative to hierarchical models [10]. This model generates both mono and stereo audio while allowing creators to condition the output on text descriptions or melodic features, which gives them even more control over the music it makes.

In conclusion, deep learning has completely changed how music is made. With techniques like Bi-LSTM, Reinforcement Learning, and Transformer Models not only help to keep traditional music styles alive, but they also make it possible for innovative musical compositions that combine human creativity with machine learning.

2 Method

2.1 LSTM-based deep learning models

Long Short-Term Memory (LSTM) models are effective in music generation because they can capture long-range dependencies in musical sequences. They are particularly useful for tasks like chord progression and melody generation, where maintaining a coherent temporal structure is crucial. LSTM networks can process sequences of musical notes or chords, ensuring smooth transitions between elements.

- Li et al. [2] applied a Bi-LSTM model combined with an Actor-Critic reinforcement learning algorithm to generate Xi'an drum music. The model used music theory-based reward systems to maintain harmony.
- Li et al. [1] used LSTM networks combined with a Hidden Markov Model (HMM) for chord-based music generation, focusing on generating multi-style chord progressions.
- Zhu et al. [8] combined an LSTM network with a Grey Wolf Optimizer (GWO) to enhance hyperparameter tuning, generating coherent music sequences that adhere to traditional musical attributes like melody and harmony.

The shared feature of these methods is the ability of LSTMs to handle the temporal dependencies in music data, allowing for accurate and stylistically consistent music generation.

2.2 Transformer-based models

Transformer models excel in music generation due to their capacity to model long-range dependencies across large token sequences. They allow for both conditional generation (based on text or musical input) and non-autoregressive generation, where entire musical sequences or stems are produced simultaneously rather than step-by-step. This makes them particularly efficient for generating high-fidelity, context-aware music. Some transformer models simplify the generation process by using token interleaving, where compressed audio tokens are processed in a single stage, enabling faster and more precise generation of both mono and stereo tracks.

- Copet et al. [10] developed MUSICGEN, a transformer-based model that uses single-stage token interleaving to process compressed audio tokens for both mono and stereo music generation. This model simplifies music generation by using a single-stage autoregressive approach.
- Parker et al. [7] introduced STEMGEN, a non-autoregressive transformer-based model designed to generate individual stems that align with the context of existing music compositions. This model focuses on iterative refinement and includes features like multi-source classifier-free guidance.

Both models showcase the transformer architecture's efficiency in handling large data streams and generating high-quality music based on detailed input.

2.3 Reinforcement learning and human feedback models

Reinforcement learning (RL) models are used in music generation to align AI outputs with human preferences. These models often involve a feedback loop where users provide real-time evaluations of the generated music. The feedback is used to refine the model, optimizing the output based on subjective qualities like musicality and emotional resonance.

- Cideron et al. [6] used RLHF in their MusicRL model, allowing the system to learn from over 300,000 user preferences. The model was fine-tuned to balance subjective musicality and text adherence, resulting in two variants: MusicRL-U and MusicRL-RU.
- Marcos et al. [3] integrated Bayesian Optimization with human feedback to adjust the parameters of their music generation model, allowing users to provide real-time feedback on the generated outputs.

Both methods emphasize the importance of human input and adaptive feedback to improve the subjective quality and relevance of AI-generated music.

2.4 Diffusion-based models

Diffusion-based models could produce high-quality audio by progressively refining random noise into structured outputs. These models typically operate by gradually "denoising" latent variables, transforming them step-by-step into coherent musical compositions. The key advantage of diffusion models is their capacity to handle both coarse and fine acoustic details simultaneously, which makes them suitable for generating complex and rich musical sequences.

- Lam et al. [11] proposed MeLoDy, a diffusion-based model guided by language models (LMs) that significantly reduces computational costs through its Dual-Path Diffusion (DPD) architecture. MeLoDy supports real-time music generation while maintaining high fidelity.
- Wu et al. [4] introduced Music ControlNet, a diffusion-based model offering time-specific control over musical elements, such as melody and rhythm, through UNet-based neural networks and mel spectrograms as intermediate representations.
- Chen et al. [9] developed MusicLDM, another diffusion-based model that uses beat-synchronous mixup strategies to improve the novelty of generated music and avoid overfitting.

These models utilize diffusion techniques to control musical attributes at a granular level, providing flexibility and precision in the generated compositions.

2.5 Emotion-conditioned music generation

Emotion-conditioned music generation models focus on aligning the generated music with specific emotional tones. These models use input features like Valence-Arousal-Dominance (VAD), which quantify emotions based on their intensity and type (e.g., happy, sad, energetic). By conditioning the music generation on these emotional parameters, the model can produce music that resonates with the intended emotional cues. This approach is particularly useful for applications requiring mood-specific compositions, such as background music for films or therapy sessions. The model's architecture typically integrates emotional features alongside text or musical input, ensuring that the output aligns with the desired emotional state.

- Sun et al. [5] introduced ECMusicLM, an extension of MusicLM, which incorporates Valence-Arousal-Dominance (VAD) emotional features to enhance the emotional depth of generated music. This model aligns the music with emotional cues from both text and musical inputs.

Emotion-conditioned models ensure that AI-generated music conveys the intended emotional impact, enhancing user experience and creative possibilities.

2.6 Music understanding and captioning models

Music understanding and captioning models are designed to interpret and describe musical content, enabling deeper interaction between AI systems and music. These models use music representations (such as embeddings or spectrograms) and pair them with pretrained language models to generate textual descriptions of music or answer music-related questions. This process involves breaking down complex musical features like rhythm, melody, and harmony into understandable components that the model can relate to text. These models are trained on paired datasets of music and descriptive text, allowing them to produce accurate captions or respond to inquiries about the structure or style of a given piece of music. This method is particularly valuable in music analysis and education, where interpreting and explaining musical content is essential.

- Liu et al. [12] developed MU-LLaMA, a model for music question answering (QA) and music captioning. By leveraging the Music Event Representation Transformer (MERT) encoder, MU-LLaMA excels in understanding and generating textual responses based on detailed music features.

These models combine advanced music comprehension with language models, allowing AI systems to not only generate music but also interact with and explain it.

3 Discussion

3.1 Advantages and disadvantages of different models

In AI-driven music generation, it is easy to foresee that each model has its own distinct advantages and disadvantages. LSTM-based models are highly effective for handling sequential data, capturing long-range dependencies, and generating smooth transitions in melodies and chord progressions. Due to their recursive nature, these models are particularly effective in the composition of simple folk and classical music, where temporal order is a critical factor. LSTMs, on the other hand, need a lot of processing power and have trouble with very long sequences, which can cause problems when trying to scale up complex musical pieces [1, 2].

On the other hand, transformer-based models bring efficiency and scalability to the forefront. These models can process large token sequences in parallel, allowing for high-quality and context-aware music generation. MUSICGEN and STEMGEN showcase transformers' ability to generate multi-instrument, long-range music pieces with high fidelity. Despite their impressive results, transformers require significant computational resources, making them less suitable for real-time applications. Additionally, the complexity of these models' training processes can be a hurdle for broader usage [7, 10].

Reinforcement learning (RL) models and those utilizing human feedback offer personalized music generation, aligning closely with user preferences. MusicRL integrates reinforcement learning with real-time feedback loops to continuously refine its outputs, producing music that resonates with human tastes. However, the reliance on subjective feedback introduces variability in results, and the feedback collection process is often resource-intensive [3, 6].

Similarly, diffusion-based models have shown potential in generating high-fidelity music by progressively refining noise into coherent musical compositions. Models like MeLoDy and MusicLDM offer granular control over musical elements like melody, harmony, and rhythm. However, the slow sampling process and computational expense make diffusion models less practical for real-time music generation [9, 11]

Emotion-conditioned music generation models, such as ECMusicLM, allow for the creation of emotionally resonant music by conditioning the generation process on emotional cues. These models excel in mood-specific music creation but struggle when the emotional input is overly dependent on textual data, limiting their flexibility [5]. Lastly, music understanding and captioning models, like MU-LLaMA, provide interpretive capabilities, generating captions or answering music-related questions. Their performance is, however, constrained by the availability of high-quality paired music-text datasets [12].

3.2 Limitations and challenges

Across all AI-driven music generation models, several overarching limitations and challenges persist. A common issue is the high computational complexity required to train and generate music, especially in real-time applications. Models like LSTM, transformers,

and diffusion-based architectures often demand significant processing power, making them resource-intensive and difficult to scale for large-scale or real-time music generation. This limits their practicality for widespread or interactive use [1, 7, 10]. Another broad challenge is data dependency. Most models require extensive, high-quality datasets to generate accurate and diverse music. However, the availability of comprehensive music datasets, particularly for non-Western or niche musical genres, remains limited [2, 12]. This constrains the models' ability to generalize across different styles and cultural contexts, often leading to less diverse outputs. Furthermore, subjectivity in evaluation poses difficulties. Models that rely on human feedback, like reinforcement learning, face inconsistencies due to the varying tastes and preferences of users. This subjectivity complicates the process of tuning models to produce music that appeals universally, as individual responses to music are highly personal and variable [3, 6]. Finally, complexity in control and user interaction is another limitation. Many models, particularly those with time-varying controls like Music ControlNet, require a level of technical expertise from users to manage detailed musical attributes. This complexity reduces accessibility, making it challenging for users without specialized knowledge to fully leverage these systems [4].

In summary, despite their potential, AI-driven music generation models are hindered by computational inefficiencies, data limitations, subjective evaluation challenges, and complex user interfaces, which all limit their widespread adoption and real-time applicability.

3.3 Future prospects

The future of AI-driven music generation holds great promise, with various avenues for advancement across multiple models. LSTM-based models could benefit from more refined harmony modeling and improved hyperparameter optimization techniques, such as integrating evolutionary algorithms. Expanding these models to accommodate a broader range of musical genres, beyond the conventional classical and pop, will enhance their versatility and relevance [1, 2].

For transformer-based models, enhancing computational efficiency is a key priority. New techniques, such as those used in diffusion models, could bridge the gap between high-quality output and real-time application, making interactive music generation more feasible [10, 11]. Incorporating these transformer-based models into professional music production workflows could offer musicians precise control over musical elements like tempo, rhythm, and instrumentation [7].

Reinforcement learning (RL) and human feedback models have the potential to become more interactive and scalable. Future models could integrate human feedback more dynamically, allowing users to influence the music generation process in real time. Moreover, developing more sophisticated reward systems that account for emotional depth, musicality, and personal preferences will improve the relevance and appeal of the generated music [6].

In diffusion-based models, improving efficiency and overcoming data limitations are essential future steps. Techniques like beat-synchronous mixup strategies, as seen in MusicLDM, could help address data scarcity while ensuring that generated music remains original and diverse [9].

For emotion-conditioned models, expanding input modalities to include physiological or visual signals could enhance the emotional resonance of generated music. This would allow for deeper emotional interaction with the user, broadening creative possibilities beyond text-based inputs [5].

Lastly, music understanding and captioning models would benefit greatly from larger and more diverse datasets, improving their ability to interpret and describe a wider range of musical styles. Expanding cross-modal training, such as integrating audio, text, and video inputs, would make these models more versatile, especially in educational and analytical

contexts [12]. Music ControlNet also holds potential for professional music production by simplifying user interfaces and offering intuitive control over time-specific musical attributes [4].

4 Conclusion

This review has mainly discussed how deep learning, through models like LSTMs, transformers, reinforcement learning, and diffusion models, is transforming music generation. Through these technologies, AI may generate music that both preserves traditional structures and introduces creative innovations. LSTM-based models are effective for sequential tasks like chord progression but struggle with scalability. On the other hand, transformer models may produce high-quality music but require a lot of computational power. Reinforcement learning and human feedback models, like MusicRL, personalize music based on user preferences, while diffusion-based models improve audio fidelity but face limitations in real-time applications. Even with these improvements, there are still challenges, such as computational inefficiencies and the need for high-quality, large-scale datasets. In the future, research should focus on improving the efficiency of transformer and diffusion models for real-time applications and integrating more diverse input modalities, such as physiological signals, to enhance emotion-based music generation.

References

1. F. Li, Chord-based music generation using long short-term memory neural networks in the context of Artificial Intelligence. J. Supercomput. 80, 6068–6092 (2023).
2. P. Li, J. Wu, X. Wang, A novel Xi'an Drum Music Generation Method based on Bi-LSTM Deep Reinforcement Learning. Appl. Intell. 54, 80–94 (2023).
3. M. Marcos, L. Jiménez, R. Ortega, GPianoroll: A Deep Learning System with Human Feedback for Music Generation. Rev. J. Jóvenes Investigadores del I3A 12, 1–10 (2024).
4. S.-L. Wu, Y. Jin, Music ControlNet: Multiple time-varying controls for music generation. IEEE/ACM Trans. Audio Speech Lang. Process. 32, 2692–2703 (2024).
5. Y. Sun, R. Huang, Emotion-conditioned musiclm: Enhancing emotional resonance in music generation. 2024 IEEE Congr. Evol. Comput. (CEC) 1, 1–8 (2024).
6. G. Cideron, E. Ricci, MusicRL: Aligning Music Generation to Human Preferences. arXiv preprint arXiv:2402.04229 (2024)
7. J.D. Parker, A. Ng, STEMGEN: A music generation model that listens. 2024 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP), 1116–1120 (2024).
8. Q. Zhu, Y. Zhang, Z. Li, Grey Wolf optimizer based deep learning mechanism for music composition with data analysis. Appl. Soft Comput. 111294 (2024).
9. K. Chen, J. Lee, MusicLDM: Enhancing Novelty in Text-to-Music Generation Using Beat-Synchronous Mixup Strategies. 2024 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP) (2024).
10. J. Copet, F. Meier, Simple and Controllable Music Generation. arXiv preprint arXiv:2306.05284 (2024)
11. M.W.Y. Lam, H. Chen, Efficient Neural Music Generation. arXiv preprint arXiv:2305.15719 (2023)
12. S. Liu, Y. Zhao, Music Understanding LLaMA: Advancing Text-to-Music Generation with Question Answering and Captioning. 2024 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP) (2024).