Deep Learning for Style Transfer and Experimentation with Audio Effects and Music Creation

Ada Tur
McGill University
Montreal, QC
ada.tur@mail.mcgill.ca

Abstract

Recent advancements in deep learning have the potential to transform the process of writing and creating music. Models that have the potential to capture and analyze higher-level representations of music and audio can serve to change the field of digital signal processing. In this statement, I propose a set of Music+AI methods that serves to assist with the writing of and melodies, modelling and transferring of timbres, applying a wide variety of audio effects, including research into experimental audio effects, and production of audio samples using style transfers. Writing and producing music is a tedious task that is notably difficult to become proficient in, as many tools to create music both cost sums money and require long-term commitments to study. An all-encompassing framework for music processing would make the process much more accessible and simple and would allow for human art to work alongside technology to advance.

Introduction and Related Work

There have been many endeavors in conducting similar experiments to those described in this proposal. Firstly, experimentation with the creation of new audio effects is an area of audio processing that has had some research conducted, though not extensively. Audio effects are methods of altering the pitch, rhythm, timbre, volume, etc of signals. Some common audio effects are distortion, reverb, and delay, though many more can be found often in music. Research into using conditional temporal convolutional networks to model audio effects has been conducted in a paper by Steinmetz and Reiss (Steinmetz and Reiss 2021). The work utilizes the model to take in audio samples paired with a target audio effect in an attempt to apply it to the sample. An interesting finding in a similar work is that the architecture of deep learning models plays a part in the kinds of audio effects that can be generated (Steinmetz and Reiss 2020). Though randomizing parameters can introduce interesting audio effects, deeper models are intrinsically linked to reverb and delay-heavy effect, whereas more shallow networks tend to produce distortion-heavy effects.

Further, research into style transfers has been conducted to learn about the capacities of neural networks in transferring the production style of an audio sample to another (Steinmetz, Bryan, and Reiss 2022). An interesting technique conducted by this work is the usage of audio effects within the computation graph of the network, such that the computer can use learned effects to assist with style transfer.

Similarly, there has been some advancement in the transferring of timbre between audio samples (Huang et al. 2018). Timbre is essentially the distinguishing factor between two instruments playing the same note with the same intensity - how can we describe why a guitar and a trumpet playing the same note sound so different? This work uses style transfer methods commonly applied to images to a time-frequency representation of audio. It then produces a waveform for the audio sample using a conditional WaveNet synthesizer. This work finds that transferring of timbre can be easily conducted using traditional methods applied to different forms of data.

An interesting progression from all of these previous works would be the incorporation of natural language processing such that a user who is not proficient in the intricacies of audio processing can easily generate and transfer audio effects using natural language descriptions (i.e. ”apply an effect that has some light delay and distortion, and a lot of reverb and flanger”). This has been studied in the context of general audio, though not extensively with music (Paissan et al. 2023).

Prior Work By Author

Throughout my research experience, I have conducted multiple experiments w Some previous work conducted by myself in the past has been into researching modelling the correlations between lyrics, melodies, and genres with audio samples. The work, which we could not publish due to copyright of lyrics data, was titled ”Genre, Artist, and Melody Dependent Lyric Generation Using Pretrained Transformer Based Language Models”. In a three-part process, a BERT model was used to classify a set of song lyrics by genre, a GPT-2 model generated new lyrics based on a set of lyrics corresponding to a genre, and then another GPT-2 model generated lyrics given MIDI samples of audio (Devlin et al. 2019), (Radford et al. 2019). This work was an introductory experiment into learning about the links a computer can have between natural language and audio samples - a technique that would be very interesting to extend for this proposal.

Approach

For this project, there are a variety of models to choose from. The first part of the project, the utilization of natural language as a form of creating new audio effects, MusicGen (Copet et al. 2023), would be efficient. Essentially, we want to use a model that is proficient in creating connections between natural language, such as "more reverb", and applying this to the audio production pipeline (resulting in model parameters that add more reverb to the sample). For the analysis of audio samples themselves, popular techniques have been VQ-VAEs, especially for audio generation (Dhariwal et al. 2020). Conditional temporal convolutional neural networks (TCNs) have also proved to be good for analyzing effects and musical intricacies within audio samples (Steinmetz and Reiss 2021). However, for this work, the MusicGen module of Meta's latest audio generation model, AudioCraft, for music generation would be the optimal choice (Copet et al. 2023). A combination of these models to work together to encompass all tasks for this environment would be the optimal approach.

Evaluation

A large challenge of working with deep learning for music is the evaluation metric - most tasks in this area require human evaluation, the basis of which can be tricky to optimize. For this task, the best evaluation method would be to train a model over some musical data, and compare various generation outputs to see if the musical data model "prefers" this output. Additionally, a human evaluation approach could be utilized. This would be done by conducting a survey, where a set of participants is given an audio sample prior to model alterations, and then the post-production version of the same sample, and they can use a set of criteria to describe the final product ("Excellent", "Some Improvement", "Worse Than Before", etc). Further, a set of musicians who are experienced with using audio software can utilize the framework and return feedback on its performance.

Discussion

The models employed are expected to make interesting connections between language, signals, and music, such that we see an example of a computer thinking in a way most human musicians would not. It would be excellent to see very interesting audio effects applied to samples and high-quality production on behalf of the toolkit. This level of progress would change the field of music, making it easier for users to write and produce music and introduce new ideas to their work. Instead of replacing the musician, it would be fantastic to see a joint-approach, with the human creating art a computer cannot, and the computer improving and enhancing this art to create distinctive and remarkable sounds. New ways of playing around with audio samples would be introduced in a way that allows for more creative freedom for artists. Music as a whole would become more accessible and available for newer musicians and artists.

Conclusion

In conclusion, music is a field that has come very far with just simple human methods and technologies. However, with an ever-changing field of computer science, artificial intelligence, and digital signal processing, it is becoming more and more possible to incorporate artificial intelligence into the process of making art. An all-encompassing music+AI toolkit that seeks to enhance audio production, music generation and writing, and experimental freedom into various audio methods and effects would progress the field of music and allow for human art to work alongside technology. A basis of quality determined by human evaluation would add to the nuanced definition of "good music and production", allowing for more interesting and creative results on behalf of a model. Music is a field that everyone deserves to partake in, and is one of the greatest creative tendencies of humans - it separates us from other living beings. New and different approaches to making and improving music will progress art as a whole and open the door for new advancements in human creation.

References

Copet, J.; Kreuk, F.; Gat, I.; Remez, T.; Kant, D.; Synnaeve, G.; Adi, Y.; and Défossez, A. 2023. Simple and Controllable Music Generation. *arXiv preprint arXiv:2306.05284*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805.

Dhariwal, P.; Jun, H.; Payne, C.; Kim, J. W.; Radford, A.; and Sutskever, I. 2020. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*.

Huang, S.; Li, Q.; Anil, C.; Bao, X.; Oore, S.; and Grosse, R. B. 2018. Timbretron: A wavenet (cyclegan (cqt (audio))) pipeline for musical timbre transfer. *arXiv preprint arXiv:1811.09620*.

Paissan, F.; Wang, Z.; Ravanelli, M.; Smaragdis, P.; and Subakan, C. 2023. Audio Editing with Non-Rigid Text Prompts. *arXiv preprint arXiv:2310.12858*.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Steinmetz, C. J.; Bryan, N. J.; and Reiss, J. D. 2022. Style transfer of audio effects with differentiable signal processing. *arXiv preprint arXiv:2207.08759*.

Steinmetz, C. J.; and Reiss, J. D. 2020. Randomized overdrive neural networks. *arXiv preprint arXiv:2010.04237*.

Steinmetz, C. J.; and Reiss, J. D. 2021. Steerable discovery of neural audio effects. arXiv:2112.02926.