Progress Report for CSS 586: Modeling Latent Patterns in Music

Alex Kyllo
akyllo@uw.edu
University of Washington
Bothell, WA, USA

ABSTRACT

This report explores recent research in modeling music with deep learning and provides a progress report on a course project to train a generative model of classical music on the MusicNet dataset.

KEYWORDS

deep learning, neural networks, music

1 INTRODUCTION

Machine learning models of music have interesting applications in music information retrieval and creative tools for musical artists and educators. Music is complex and challenging to model because it exhibits a hierarchy of recurring patterns.

Depending on the task, machine learning models of music may be trained on the audio signal itself, either in a time domain or a frequency domain representation, or they may be trained on a digital symbolic representation of music, the most common of which is MIDI (Musical Instrument Digital Interface) notation. MIDI is an encoding of music as streams of bytes in one or more tracks or channels, each representing a sequence of 128 possible pitch values, along with timing, pressure and instrument values. A music transcription model may convert an audio signal into MIDI, which can easily be converted into other symbolic representations such as sheet music for human performers to read from, while a synthesizer model can convert MIDI representations into audio signals.

2 RELATED WORK

Google's Magenta is an umbrella project for music deep learning research and development of software tools to expose these models for use by creative artists and students.

MusicVAE is a variational LSTM autoencoder for MIDI that incorporates a novel hierarchical structure using a "composer" recurrent layer in its encoder model to better capture structure at multiple levels [7].

Music Transformer is a generative model that borrows its approach from the Natural Language Processing (NLP) domain, using an attention network to model MIDI music as a sequence of discrete tokens with relative positional dependencies [3].

A major advantage of working with the symbolic representation of music is that it is of far lower dimensionality than the raw audio waveforms of a recorded performance, which makes it less computationally expensive. However, there are many aspects of musical performance that are not captured by a symbolic representation, so the expressiveness of symbolic generative models is constrained [4].

Other research has focused on modeling raw audio waveforms directly. WaveNet is a causal convolutional neural network for generating raw audio waveforms, developed by Google DeepMind, which achieves state of the art performance in generating natural sounding speech from text, but is also capable of generating short, realistic snippets of audio music [6].

Another model named SampleRNN generates raw audio waveforms using a three-tier hierarchy of gated recurrent units (GRU) to model recurrent structure at multiple temporal resolutions [5].

Jukebox by OpenAI utilizes a vector-quantized variational autoencoder (VQ-VAE) to compress raw audio into a sequence of discrete codes and models these sequences using autoregressive transformers to generate music [1].

Audio data can also be modeled in the frequency domain through the use of Fourier analysis. The recent MelNet model is trained on spectrograms and can learn musical structures such as melody and harmony and variations in volume, timbre and rhythm [8].

Prior work points out that the division between symbolic music notes and the sounds of music is analogous to the division between symbolic language and utterances in speech, which may inspire ideas for combining the two approaches [2]. A paper from Boston University describes an effort to combine the symbolic and waveform approaches to music modeling, by training an LSTM to learn melodic structure of different styles of music, then providing generations from this model as conditioning inputs to a WaveNet-based raw audio generator [4].

3 PLANNED METHODS AND PROGRESS

3.1 Data Preprocessing

This project will focus on the generative modeling of symbolic music using MIDI data, because of the advantages of symbolic music models in representing long-term structure in musical compositions to produce generations with coherent use of repetition over long time scales.

3.2 Model Fitting

We will explore several modeling approaches to generating symbolic music:

- Sliding window sequence prediction with RNNs (LSTM/GRU)
- Sliding window sequence prediction with Transformers
- Latent space interpolation with Sequential VAEs
- Latent space interpolation with Sequential GANs

3.3 Model Evaluation

Evaluation of generative models is challenging because there is no equivalent of an accuracy metric like what is used in supervised learning. Generative models are typically evaluated using a combination of qualitative metrics whereby human judges rate the quality of the generated examples (essentially a Turing test), and quantitative metrics that assess the differences in the parametric distributions of generated and real examples. Yang and Lerch (2020) proposes a set of metrics informed by music theory, for probabilistically evaluating how similar the generations are to known sample distributions of real music [9]. These metrics include counts, ranges, histograms and transition matrices of pitches and note lengths, then the Kullback-Leibler divergence and overlapping area of the probability density functions are use to compare against known reference distributions per musical genre [9]. Due to the cost and time requirements associated with designing a human subjects experiment, we plan to utilize this quantitative approach to generation quality assessment.

4 FUTURE WORK

REFERENCES

[1] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. 2020. Jukebox: A Generative Model for Music. arXiv:2005.00341 [eess.AS]

[2] Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse Engel, and Douglas Eck. 2019. Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset. arXiv:1810.12247 [cs.SD]

[3] Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M. Dai, Matthew D. Hoffman, Monica Dinculescu, and Douglas Eck. 2018. Music Transformer. (2018). arXiv:1809.04281 http://arxiv.org/abs/1809.04281

[4] Rachel Manzelli, Vijay Thakkar, Ali Siahkamari, and Brian Kulis. 2018. Conditioning Deep Generative Raw Audio Models for Structured Automatic Music. *CoRR* abs/1806.09905 (2018). arXiv:1806.09905 http://arxiv.org/abs/1806.09905

[5] Soroush Mehri, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, Jose Sotelo, Aaron Courville, and Yoshua Bengio. [n.d.]. SampleRNN: An Unconditional End-to-End Neural Audio Generation Model. ([n. d.]). arXiv:1612.07837 http://arxiv.org/abs/1612.07837

[6] Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. [n.d.]. WaveNet: A Generative Model for Raw Audio. ([n. d.]). arXiv:1609.03499 http://arxiv.org/abs/1609.03499

[7] Adam Roberts, Jesse Engel, Colin Raffel, Curtis Hawthorne, and Douglas Eck. [n.d.]. A Hierarchical Latent Vector Model for Learning Long-Term Structure in Music. In *Proceedings of the35th International Conference on Machine Learning* (2018). 10.

[8] Sean Vasquez and Mike Lewis. 2019. MelNet: A Generative Model for Audio in the Frequency Domain. arXiv:1906.01083 [eess.AS]

[9] Li-Chia Yang and Alexander Lerch. [n.d.]. On the evaluation of generative models in music. 32, 9 ([n. d.]), 4773–4784. https://doi.org/10.1007/s00521-018-3849-7