Music Generation

Generative Artificial Intelligence (GAI) is not a very well-covered area of data science at the moment, in mid-2023. For my final CS770-DL project, I intend to implement a music generation currently focused on RNNs, but not excluding LSTM, VAE, or transformer-based architecture if I have the time to delve deeper into the music generation topic, to which I must admit I am a complete novice. By contrast, my wife is a professional singer who agreed to top-line an acceptable music GAI composition I create.

I plan to focus on Musical Instrument Digital Interface (MIDI), defined as “MIDI is a technical standard that describes a communications protocol, digital interface, and electrical connectors that connect a wide variety of electronic musical instruments, computers, and related audio devices for playing, editing, and recording music.” (General MIDI ("GM") Specifications, n.d.)

Literature Review

I reviewed a significant number of papers, four of which are summarized in this section in order of their complexity.

The first, Muzika! (Marco Pasini, 2022) relies on first learning a compact invertible representation of spectrogram magnitudes and phases with adversarial autoencoders, then training a Generative Adversarial Network (GAN) on this representation for a particular music domain. A latent coordinate system enables generating arbitrarily long sequences of excerpts in parallel, while a global context vector allows the music to remain stylistically coherent through time. The authors release the source code and pretrained autoencoder weights at this http URL, such that a GAN can be trained on a new music domain with a single GPU in a matter of hours.

The second, MusicGen (Generation, 2023) tackles the task of conditional music generation as a single Language Model (LM) that operates over several streams of compressed discrete music representation, i.e., tokens. MusicGen is comprised of a single-stage transformer LM together with efficient token interleaving patterns, which eliminates the need for cascading several models, e.g., hierarchically or upsampling. MusicGen generates high-quality samples, while conditioned on textual description or melodic features, allowing better controls over the generated output. Music samples, code, and models are available at this https URL.

The third, MidiNet (Li-ChiaYang, 2017) unlike most existing music generation RNNs, the DeepMind’s WaveNet model showed that CNNs can also generate realistic music. Following this light, the authors investigate using CNNs for generating melody (a series of MIDI notes) one bar after another in the symbolic domain. In addition to the generator, they use a discriminator to learn the distributions of melodies, making it a GAN. They introduce a novel conditional mechanism that exploits available prior knowledge so that the model can generate melodies either from scratch, by following a chord sequence, or by conditioning on the melody of previous bars (e.g. a priming melody), among other possibilities. The MidiNet model, can be expanded to generate music with multiple MIDI channels (i.e. tracks). The authors’ subjective user study shows that MidiNet performs comparably with Google's MelodyRNN models as realistic and pleasant to listen to, yet MidiNet's melodies are reported to be much more interesting.

The fourth and last, MuseGAN (Hao-Wen Dong, 2017) seems to be the most sophisticated. First, music is an art of time, necessitating a temporal model. Second, music is usually composed of multiple instruments/tracks with their own temporal dynamics, but collectively they unfold over time interdependently. Lastly, musical notes are often grouped into chords, arpeggios, or melodies in polyphonic music, and thereby introducing a chronological ordering of notes is not naturally suitable. In MuseGAN, there are not two but three models for symbolic multi-track music generation under the framework of GANs -- the jamming model, the composer model, and the hybrid model that can generate coherent music of four bars right from scratch (i.e. without human inputs). The models can extend to human-AI cooperative music generation: given a specific track composed by humans, they can generate four additional tracks to accompany it. All code, the dataset, and the rendered audio samples are available at this URL.

Sample architecture image

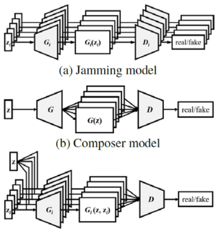

Figure 1

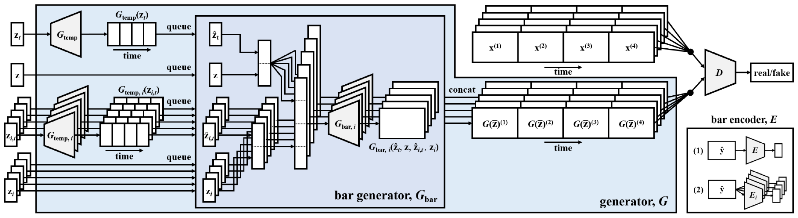

Figure 2

Figure 1 - Three GAN models for generating multi-track data. Note that we do not show the real data x, which will also be fed to the discriminator(s).

Figure 2 - System diagram of the proposed MuseGAN model for multi-track sequential data generation.

The MuseGAN (Hao-Wen Dong, 2017) seems to be the most quoted & implemented music generation model; however, it is a difficult model to implement in a limited time for this course.

Unrelated, in addition to music generation, I am learning in parallel how to use TensorBoard so in the final project submission I can attach my own architecture image professionally generated.

Dataset

The dataset I am thinking of using for the project is the million-song dataset. (Bertin-Mahieux, Ellis, Whitman, & Lamere, 2011) Alternatively, I may use the Lakh MIDI dataset (Raffel).

Data Dictionary

Music can be simplified to the following quadruple:

1. Instrument -- plays notes, whereby each note consists of:
2. Pitch -- the perceptual quality of the sound as a MIDI note number.
3. Step -- the time elapsed from the previous note or start of the track.
4. Duration -- how long the note will be playing in seconds and = note end - start times.

I will utilize 1 instrument; thus, the triplet of Pitch, Step, and Duration uniquely identifies each and every note to be generated.

Bibliography

Bertin-Mahieux, T., Ellis, D. P., Whitman, B., & Lamere, P. (2011). *THE MILLION SONG DATASET*. Retrieved 7 13, 2023, from http://tbertinmahieux.com/papers/ismir11.pdf

General MIDI ("GM") Specifications. (n.d.). Retrieved 7 13, 2023, from http://www.midi.org/techspecs/gm.php

Generation, S. a. (2023, 06 08). *Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, Alexandre Défossez.* Retrieved from arxiv: https://arxiv.org/abs/2306.05284

Hao-Wen Dong, W.-Y. H.-C.-H. (2017, 11 24). *MuseGAN: Multi-track Sequential Generative Adversarial Networks for Symbolic Music Generation and Accompaniment.* Retrieved from arxiv: https://arxiv.org/abs/1709.06298

Li-ChiaYang, S.-Y.-H. (2017, 03 31). *MIDINETACONVOLUTIONALGENERATIVEADVERSARIAL NETWORKFORSYMBOLIC-DOMAINMUSICGENERATION.* Retrieved from Papers with Code: https://paperswithcode.com/paper/midinet-a-convolutional-generative

Marco Pasini, J. S. (2022, 08 18). *Musika! Fast Infinite Waveform Music Generation.* Retrieved from arxiv: https://arxiv.org/abs/2208.08706

Raffel, C. (n.d.). *The Lakh MIDI Dataset v0.1.* Retrieved from Colin Raffel: https://colinraffel.com/projects/lmd/