Music Controllable Diffusion

Saravana Rathinam*

Stanford University
saravanr@stanford.edu

Abstract

Diffusion based generative models can be used to generate music in a controllable way. Generating long music sequences from raw audio waveforms can be very compute intensive. MIDI format gives a much more compressed representation of a subset of Music which may give us a tractable way to generate melodies.

1 Motivation

Composing music is a skill that is acquired by many years of practice. The music itself is the result of the life experiences of the musician, their state of mind, their unconcious and concious thoughts. Their creative talent is subjective and difficult to generalize. Recent advances in Generative models for Music generation have shown impressive results where the focus has been to replace the creative process. Learning the distribution of music creation may be an intractable problem at the moment. However one approach we can take is to build tools that serve as an aid in the creative process. If a musician already has a few ideas in mind on how a song or melody should start, can the problem be modelled as a conditional generative process where given the start and style of the song, can a model generate multiple possibilities of how the song can proceed?

In such a generative system, the inputs to the model would be a short MIDI sequence. The system would generate a bunch of sequences that may serve as suggested next sequences and so on. By conditioning on the input and letting the musician choose the path to take, the model can help in the creative process.

2 Related Works

There has been a lot of good work in the area of Music creation. Google's Magenta project explores the role of machine learning in the creative process. In a recent papers (Mittal et al., 2021) built a multi-stage non autoregressive generative model that enabled using diffusion models on discrete data. They generated both unconditional music as well as conditional in-filling. They used a Denoising Diffusion Probabilistic Model Ho et al. (2020) on top of a MusicVAE model that generated the continuous time latent embeddings. Similarly Choi et al. (2021) proposed an Iterative Latent Variable Refinement (ILVR) method to guide the DDPM to generate high quality images based on a given reference image. Also Song et al. (2020) produced a way to accelerate sampling process of a DDPM which can make generation process of sequences faster. In another beautiful apprach, Bazin et al. (2021) built an interactive web interface that transforms sound by inpainting. This approach is similar to what Meng et al. (2021) built with SDEdit that adapts to editing tasks at test time, without the need for re-training the model.

*Project for course CS236 - Generative Models

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

3 Dataset

The dataset for the project a combination of the Lakh MIDI Dataset v0.1 Raffel and the MIDI dataset posted at (midi man). The Lakh MIDI data set is a collection of 176, 581 unique MIDI files out of which 45, 129 have equivalent songs in the Million Song Dataset. The (midi man) collection has about 150, 000 midi files. All the MIDI files were converted to the OctupleMIDI encoding format as proposed and implemented in Zeng et al. (2021). The total number of music samples were about 290, 000.

4 Methodology

Initially the plan is to build a unconditional MusicVAE and see its performance on the LMD dataset. Following which I plan to adopt a similar approach to (Mittal et al., 2021) to build a DDPM model that has the ability to inpaint music in a manner similar to SDEdit Meng et al. (2021).

5 Evaluation

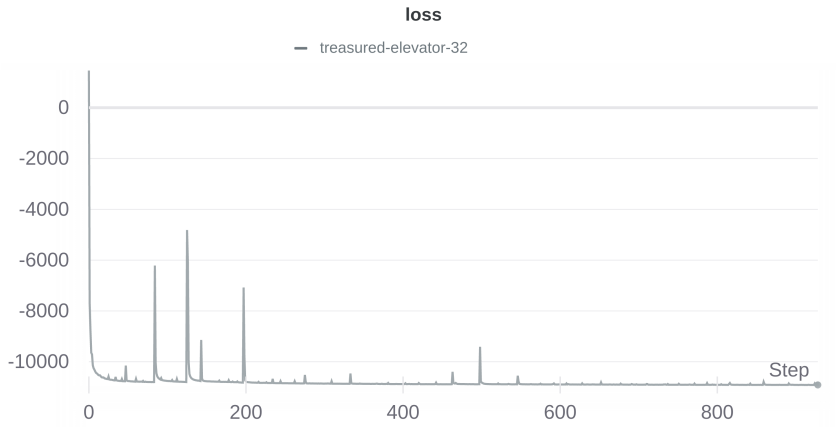
The music evaluation is a qualitative process however in related works authors have used Fréchet distance (FD) Heusel et al. (2018) and Maximum Mean Discrepancy (MMD) Gretton et al. (2012) to measure distance between the models continuous output distribution and the original data distribution in latent space.

6 Technical Approach

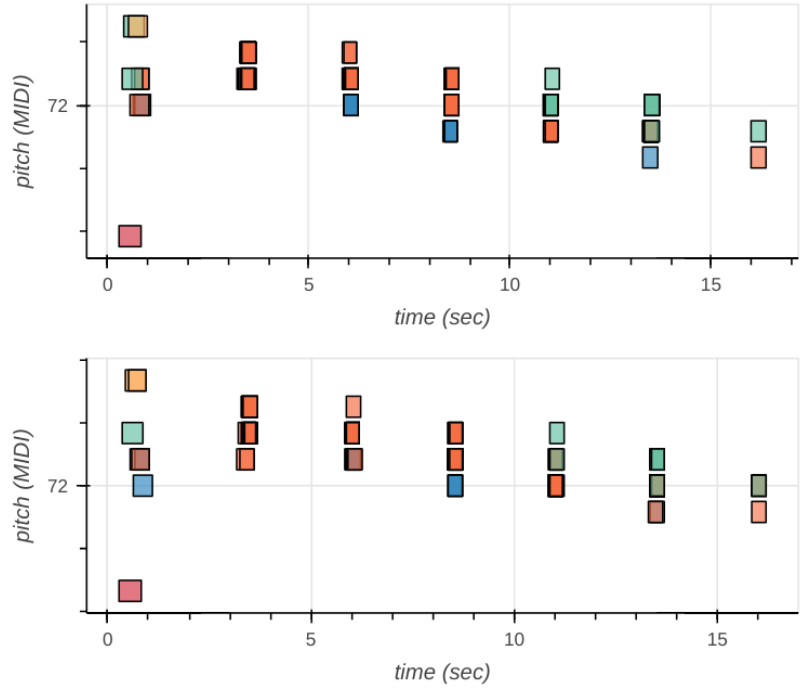
The first baseline was established using a Variational Auto Encoder.

7 Prelimnary Results

Training and Test Loss of VAE:



Pitch plot of generated MIDI files.



References

Bazin, T., Hadjeres, G., Esling, P., and Malt, M. 2021. doi: 10.30746/978-91-519-5560-5. URL http://dx.doi.org/10.30746/978-91-519-5560-5.

Choi, J., Kim, S., Jeong, Y., Gwon, Y., and Yoon, S. Ilvr: Conditioning method for denoising diffusion probabilistic models, 2021.

Gretton, A., K, K. M. B., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test, 2012. URL "http://jmlr.org/papers/v13/gretton12a.html".

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2018.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models, 2020.

Meng, C., Song, Y., Song, J., Wu, J., Zhu, J.-Y., and Ermon, S. Sdedit: Image synthesis and editing with stochastic differential equations, 2021.

midi man. The largest midi collection on the internet. URL https://www.reddit.com/r/WeAreTheMusicMakers/comments/3ajwe4/the_largest_midi_collection_on_the_internet/.

Mittal, G., Engel, J. H., Hawthorne, C., and Simon, I. Symbolic music generation with diffusion models. *CoRR*, abs/2103.16091, 2021. URL https://arxiv.org/abs/2103.16091.

Raffel, C. Motion sensors | andriod development. URL https://colinraffel.com/projects/lmd/.

Song, J., Meng, C., and Ermon, S. Denoising diffusion implicit models, 2020.

Zeng, M., Tan, X., Wang, R., Ju, Z., Qin, T., and Liu, T.-Y. Musicbert: Symbolic music understanding with large-scale pre-training, 2021.