Music Controllable Diffusion
Generating MIDI Music using Controllable Diffusion

Name: Saravana Kumar Rathinam
SUNet ID: 06512865

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.

Such a tool can help naive music enthusiasts to try creating music. Consider the spectrum of music creation tools, on one end are very sophisticated tools like Abelton Live, FL Studio which are used by trained musicians. Then there are tools in the middle like Garage band for casual users. With conditional music generation we can build a tool that lies at the other end of the spectrum, enabling anybody to try music creation.

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.) 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.

- Submissions must be in PDF.
- Submitted papers can be up to eight pages long, not including references, plus unlimited space for references. Accepted papers can be up to nine pages long, not including references, to allow authors to address reviewer comments. Any paper exceeding this length will automatically be rejected.
- **Do not include author information or acknowledgements** in your initial submission.
- Your paper should be in **10 point Times font**.
- Make sure your PDF file only uses Type-1 fonts.
- Place figure captions *under* the figure (and omit titles from inside the graphic file itself). Place table captions *over* the table.

- References must include page numbers whenever possible and be as complete as possible. Place multiple citations in chronological order.

- Do not alter the style template; in particular, do not compress the paper format by reducing the vertical spaces.

- Keep your abstract brief and self-contained, one paragraph and roughly 4–6 sentences. Gross violations will require correction at the camera-ready phase. The title should have content words capitalized.

References

Bazin, T., Hadjeres, G., Esling, P., and Malt, M. 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.

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.

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.

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

A. Do *not* have an appendix here

Do not put content after the references. Put anything that you might normally include after the references in a separate supplementary file.

We recommend that you build supplementary material in a separate document. If you must create one PDF and cut it up, please be careful to use a tool that doesn't alter the margins, and that doesn't aggressively rewrite the PDF file. pdftk usually works fine.

Please do not use Apple's preview to cut off supplementary material. In previous years it has altered margins, and created headaches at the camera-ready stage.