Exploring Generations Models for Music

Christopher Ziko, Rakin Munim

Introduction

Music is an important component of human life. Given the recent success of generative models in natural language and imagery, it has been reasonable to start exploring generative music. In this project we set out to learn about current state-of-the-art models of music generation and conduct some experiments by changing fine tuned pre-existing large models. We particularly seek to apply a diffusion based model on tokenized .midi data. We utilize the Lakh dataset as our data corpus following AMT. This poster will give you an understanding of the field of music generation and present some of our results. Keep in mind that this poster is a tiny reflection of the full field of generative music.

Motivation

Imagine yourself to be an expert musician. You can play various types of instruments, you understand how to harmonize correctly, and most importantly when you have a small idea you can quickly turn it into a song. Unfortunately, for most people this is just not possible. Music composition is a skill requiring lots of time and experience, not something everyone can quickly just do. However humans in general listen to lots of music. Most of us have an understanding of styles and themes. It would be powerful to give both non-experts and experts the ability to create music similar to what exists with using natural language. It is analogous to how computers have people who may not have the best motor control to be able to write in a readable fonts opening the door for more human communication.

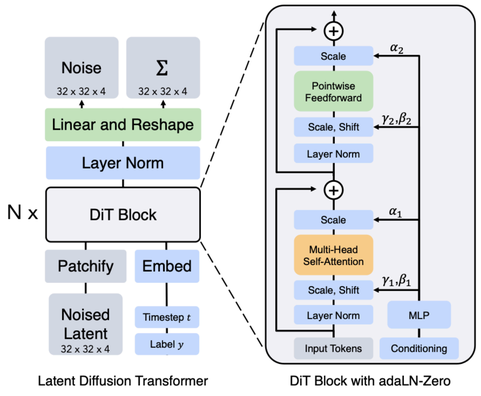


Figure 1: DiT architecture design. The model takes latent variables and predicts the noise and variance schedule

Research Objectives

- RO_1: Abstract: Build a tool that allows non-experts of music, to be able to create music in a cheap and time minimizing way.
- RO_2: Understand current SOTA, understand model abilities, limitation, and further explorations
- RO_3: Explore design patterns that are possible, even without pre-trained models
- RO_3: Fine tune an DiT on tokenized musical data

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \quad q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1})$$

$$p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T)\prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) \quad p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t))$$

Figure 2: q maps the forward noising process, whereas p maps the reverse | Source: lilianweng.github.io

Data

- There are two primary types of music storage: MP3-like Files and MIDI files. MP3 post-synthesized files containing wave data. This file can be given to a speaker and play out. MIDI files are raw "scores" giving us the raw information to synthesize the music. Since the file is raw it can be modified to change up the music at the granular level(e.g. change a note velocity), something not possible with the .mp3 file. The project will focus on MIDI based generation. MIDI files encode all data needed to reproduce a song. In a file you will find tracks, note data, pitch, velocity, etc. The Lakh Midi Dataset(LMD): The LMD dataset is a large collection of about 100K midi files. Midi files contain data(notes, keys, etc) and metadata(lyrical events, annotations) about the music. We use a subset of the dataset: clean midi. This subset contains roughly 17K songs. Here are some useful figures:

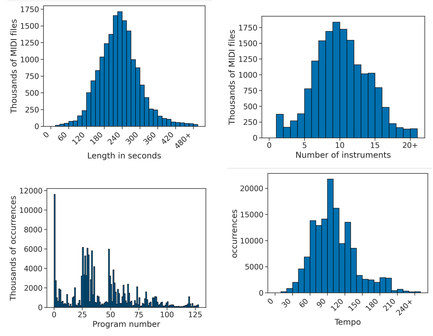


Figure 3: TL: Cadinality over length. TRcardinality over instrumets; BL: Cardinality over program number BR: Cardinality over tempos

Arch

- Language modeling is just some prediction task conditioned on some input. 2 primary models of music generation: autoregressive (AR) models and diffusion models. AR models predict tokens in a sequential scan with some fixed window size. This works well for textual data since textual data is generally autoregressive in the human form. However maybe this isn't the best approach for other forms of art. Enter diffusion based models.

Diffusion based model have been widely successful in the image generation. The diffusion architectures makes use of noise in its learning system. They way diffusion systems learn is as follows: DMs take raw inputs such as an image and continually add gaussian noise for some time steps T. At the end of the time steps the image should just look no different from random noise. This adding of noise is the forward process. The goal however is to go in the reverse direction, termed "denoising". If we can create a prediction system that given (t + 1)th state can predict t, then we can actually take random noise and go back to a photo from the true distribution.

Anticipatory Music Transformer (AMT) is a autoregressive music generative model. It utilizes the Lakh dataset and can be used for tasks such as infilling and continuation. The author's create sequences for tokenization, instead of a pictoral approach the DiT takes. The main contribution of the model is that it can be use to generate accompaniments conditioned upon pre-given melodies. This allows for a special technique, named 'anticipation' by the authors. Instead of a typical autoregressive model just taking in a fixed set of token and then auto regressing all future tokens, the model interweaves a set of control tokens as the sequence gets generated.

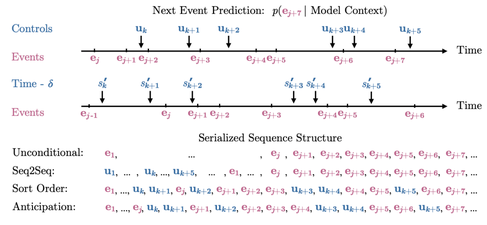
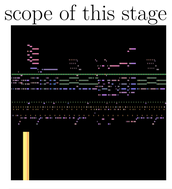


Figure 4: A visualization of anticipation in the sequence

Experimental Design & Results

For our experiment we will make use of Diffusion Transformer by Facebook. Under the hood DiT utilizes a VAE, which has a downsampling rate of 8 (256 x 256 x 3) -> (32 x 32 x 4). Latentization is an important component in making diffusion models scale well. Additionally, DiT also has the ability to have conditional label, enabling guided generation. This is an important part of future experiments with diffusion models. Remember AMT has the ability to auto-regress with anticipation. Hence it is a goal with our chosen model to have some mechanism through which we can enable anticipation. For our project however our main object is to train with a new final layer head, image based musical representation.

However there is a problem with diffusion models. MIDI files come as sequential data. To make use of MIDI files in the DiT we must first find a way to vectorize the songs so that it is image-like. This is typically done using something called piano rolls. These are image-like. We employ off the shelf midi-to-image systems to create colored piano rolls. We seek to explore two tokenizers, midipic and midi-to-images, to do the conversion. These tokenizers are lossy, however for this stage of the project, we accept this to focus on diffusion. This is an area of research that is out of the scope of this stage of the project.



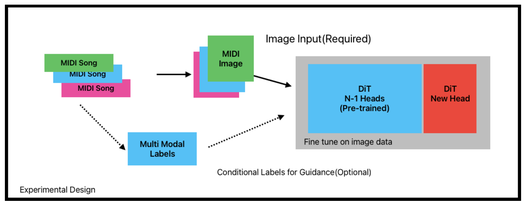


Figure 5: Top: Dancing queen as a colored piano rolled. Created using the midipic library. Bottom: Our experimental design

Results: Find out on presentation day!

Conclusion

In this project we scope out the field of generative music. We map out the space of both autoregressive models and diffusion based models. We present a experiment that makes use of a pre-trained diffusion transformer and fine-tune it on a subset of the Lakh MIDI dataset.