BbP: Milestone 2

Anonymous ACL submission

Abstract

In this second milestone we discuss the data type we will use and the metrics we will be using to evaluate our model's performance. We also share updates on our goal for the project. *This paper is still riddled withe errors*

1 Body

1.1 Pivot

We are making a modification to our plans for research and end goal. Initially we had planned to work on lyrical data, but we have pivoted to working with MIDI data. (This format will be discussed in the 'Data' subsection). Our goal remains to change the architecture from an auto regressive model to a diffusion based systems and report the evaluation metrics.

1.2 Data

The musical data corpus is quite big. Classical composers have written "scores" for centuries. Additionally with modern instruments such as MIDI-keyboards, the data that exists for music is quite large. Musical data is stored in two main ways. The first can be understood using the popular file format '.mp3'. An mp3 file contains the information a speaker needs to play that audio. However it does not provide the underlying notes that were used to create that file. If a keen human were to listen to this audio file they might be able to pick out some of the notes, octaves, etc.

The other way musical information can be encoded is through the a static object that doesn't necessarily allow you to hear the music, but provides all the underlying notes. This is essentially what the score is. This score can be given to a synthesizer which then can "play" the song for us to hear. But it can also be experienced without being played. Many musicians spend time trying to understand some of the underlying patterns of successful music.

In our research we will be using the second format. There are many abstractions allowing musical score data to be stored on a computer. The particular type we will use is the MIDI format. MIDI is an interface that allows computers to speak to musical keyboards. This file format can be used to construct various representation of musical scores. One of these is the humdrum representation.

The paper we aim to model our research over used the Lakh MIDI Dataset v0.1. **?** This dataset has been through cleaning steps such as de-duplication. Simple statistics that come with the data suggests that it is quite diverse. The dataset contains more than 150,000 songs. Most of the songs are longer than 60 seconds, with the average being around 200 seconds. The data also makes use of a diverse set of instruments giving us the ability to have our model generalize to a wide range of instruments. The average number of instruments falls around 10.

1.3 Metrics

The papers by Thickstun and Rütte both mention various metric for performance evaluation. However perplexity is used by both papers as the main metric of fluency. Rütte specifically writes about the reasoning behind the choice being that, "perplexity measures the likelihood of sequences while normalizing over the sequence length, which makes it better suited to comparing sequences of different lengths than the negative log-likelihood"**?**.

Important metrics other than perplexity there also are cited by the paper such as cosine similarity and macro-overlapping area. Additionally, Rütte compares fidelity and accurancy of the songs produces. These are both important metrics that contain information that perplexity might miss. For example a fidelity metric may reflect the amount of notes being played over a standard unit of time. We can call this the density. Thus when comparing models we should take this fidelity measure

in to provide a more complete picture. In our experiments we plan on using the metrics outlined in Rütte's paper to understand our models performance. We also will try to factor in the model's cost. This includes training time, dataset size, compute resources, etc.

1.4 Baseline

For our baseline we will adopt the two baselines. We will compare our work to Thickstun's performance metrics and we will also compare to their baseline, which is the FIGARO music transformer mentioned before. The FIGARO music transformer is a description-to-sequence model, which means instead of taking in a set of control notes, as the AMT did, they taken in a textual description of the music the user seeks to listen to and the system tries to output a matching midi file. The FIGARO also makes use of the Lakh MIDI dataset hence comparing to both Thickstun and Rütte's papers easy. Since we have not altered the underlying datasets, we will simple use these baseline figures as reported. In future milestone we will provide a more complete table of this data, but for now here is a simple description. Rütte's model achieved a PPL score of around 1.78. Thinkstun reported four different types of PPL, but the lowest was 1.52 which was on their largest model. We hope to compare: fluency, fidelity, accuracy, and cost metrics between these three models. (2, + 1(ours))