CS5647 Project Proposal:
Piano Music Generation from Text Description

Introduction

With the emergence and advent development of Large-Language-Models (LLMs), text to music generation has undergone a revolutionary transformation, opening up an unprecedented prospect for innovation, creativity, and accessibility in music composition. For instance, MusicLM [1], an experimental tool from Google that turns written prompts into music, focuses on dealing with general descriptive phrases such as “soft music that is easy for beginners” based on a pretrained joint music-text model called MuLan [2], allowing beginners to be engaged in music composition. Different from MusicLM, MuseCoco [3] proposed by Microsoft, is to generate symbolic music based on text descriptions with precise musical attributes, aiming to increase the efficiency and editability in creating music from scratch for skilled musicians.

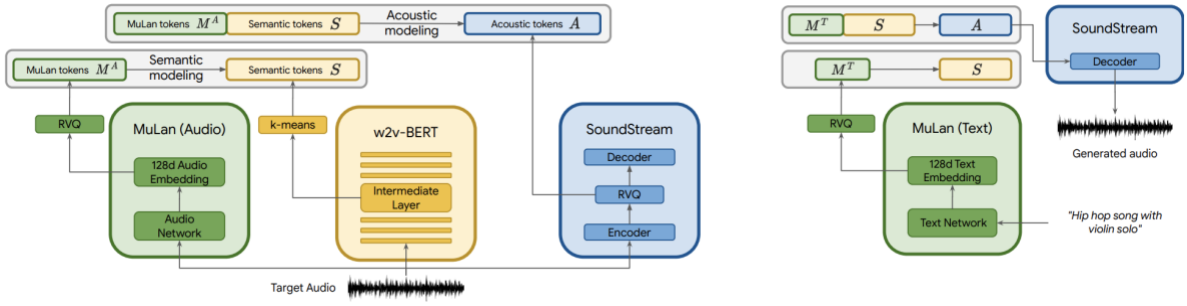
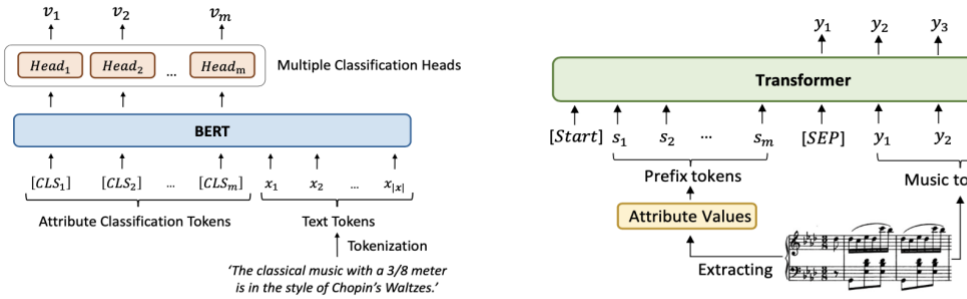


Figure 1. Training Process in MusicLM



(a) Text-to-attribute understanding (b) Attribute-to-music generation

Figure 2. Two-stage Training Pipeline in MuseCoco

However, the research topic on piano music generation from textual descriptions is less explored and has seen slower process, mainly due to the scarcity of the paired text-music data for piano music generation.

In this project, we want to leverage LLMs to develop a controllable automatic piano music generation tool that can generate music with both precise and general text descriptions, bridging the gap between these two types of text descriptions. By allowing users to define precise musical attributes including pitch, tempo, key, and chord progression, our objective is to help the pianists or composers can transform their initial inspiration in mind into an actual melody easily. Moreover, we hope this tool can pave the way for both skilled musicians and

beginners to engage in creative experimentation, exploring diverse combinations of these fundamental musical elements with ease.

Objective

The primary objective of this project is to develop a text-to-music generation system based on LLMs that can seamlessly transition between general and precise textual descriptions, providing musicians with bimodal flexibility, both in terms of compositional freedom and fine-grained control over musical attributes, thus enriching the creative process while ensuring the highest possible quality and coherence of each musical composition.

Data Collection and Preparation

Our data came from the MAESTRO [4] dataset, a collection consists of about 200 hours of paired audio and MIDI recordings from ten years of International Piano-e-Competition. The MIDI data includes key strike velocities and sustain/sostenuto/una corda pedal positions. Audio and MIDI files are aligned with ~3 ms accuracy and sliced to individual musical pieces, which are annotated with composer, title, and year of performance. We used v2.0.0 version of the dataset which provides preserved sostenuto (CC 66) and una corda (CC 67) messages in MIDI files with a new train/validation/test split.

Methodology

Since we have collected a well-annotated dataset which contains paired piano audio and MIDI recordings annotated with basic information including compose, title and year of performance and etc, our work mainly focuses on the following key aspects:

1. Extraction of Musical Attributes from MIDI File: For each MIDI file, extract the same list of musical attributes.
2. Textual Descriptions Generation Through Prompt Learning: Employ prompt learning to generate a general descriptive text for each piano audio.
3. Creation of a Well-Annotated Dataset: Combine the labels generated in previous two steps, curate an annotated dataset comprising <MIDI, text> pairs and each MIDI file is labelled with the same list of musical attributes it has.
4. Text Encoding with LLM-Based Encoder: Design and train a text encoder by comparing the performance and adaptabilities of different LLMs. This encoder takes textual descriptions as input, producing corresponding text embeddings that streamline subsequent stages of our music generation pipeline.
5. Conversion of MIDI into Sequential and Discrete tokens Using REMI Event Representation [5]: Transform MIDI file into a sequence of text-like discrete tokens using REMI as the event representation.
6. Development of the Piano Music Generation Model.
7. Further Experiments and Evaluations on Performance.

Reference

[1] A. Agostinelli et al., "MusicLM: Generating Music From Text," arXiv:2301.11325 [cs.SD], 2023.
[2] Q. Huang et al., "MuLan: A Joint Embedding of Music Audio and Natural Language," arXiv:2208.12415 [eess.AS], 2022.
[2] P. Lu et al., "MuseCoco: Generating Symbolic Music from Text," arXiv:2306.00110 [cs.SD], 2023.
[4] C. Hawthorne et al., "Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset," in Proceedings of the International Conference on Learning Representations, 2019, available at: https://openreview.net/forum?id=r1lYRjC9F7.
[5] Y. Huang and Y. Yang, "Pop Music Transformer: Beat-Based Modeling and Generation of Expressive Pop Piano Compositions," in Proceedings of the ACM Multimedia Conference (MM '20), Seattle, WA, USA, 2020, pp. 1180-1188, doi: 10.1145/3394171.3413671.