related work for the project

April 7, 2023

Contents

1 Code

1.1 hugging face diffusers

- `https://github.com/huggingface/diffusers`
- Diffusers is the go-to library for state-of-the-art pretrained diffusion models for generating images, audio, and even 3D structures of molecules.

1.2 audio-diffusion-pytorch (ETH + max-planck)

- `https://github.com/archinetai/audio-diffusion-pytorch`
- A fully featured audio diffusion library, for PyTorch. Includes models for unconditional audio generation, text-conditional audio generation, diffusion autoencoding, upsampling, and vocoding. The provided models are waveform-based, however, the U-Net (built using a-unet), DiffusionModel, diffusion method, and diffusion samplers are both generic to any dimension and highly customizable to work on other formats. Note: no pre-trained models are provided here, this library is meant for research purposes.

1.3 dance diffusion (harmonai)

- `https://colab.research.google.com/github/Harmonai-org/sample-generator/blob/main/Dance_Diffusion.ipynb?pli=1#scrollTo=lU97ZiP7nSKS`
- Unconditional random audio sample generation
- Audio sample regeneration/style transfer using a single audio file or recording
- Audio interpolation between two audio files

1.4 TODO diffusion lm and controllable music gen

- https://github.com/SwordElucidator/Diffusion-LM-on-Symbolic-Music-Generation

1.5 TODO FIGARO: Controllable Music Generation using Learned and Expert Features

- multitrack symbolic multi-track midi generator with expert-defined control features
- https://colab.research.google.com/drive/1UAKFkbPQTfkYMq1GxXfGZOJXOXU_svo6#scrollTo=Zsszt4J46OIj

2 Literature

2.1 FIGARO: Generating Symbolic Music with Fine-Grained Artistic Control

- https://arxiv.org/pdf/2201.10936.pdf

2.2 FIGARO: CONTROLLABLE MUSIC GENERATION USING EXPERT AND LEARNED FEATURES

- https://openreview.net/pdf/4ee95daea73cb05d2ea8780258b25684ccd82a88.pdf (more recent)
- Recent symbolic music generative models have achieved significant improvements

in the quality of the generated samples. Nevertheless, it remains hard for users to control the output in such a way that it matches their expectation. To address this limitation, high-level, human-interpretable conditioning is essential. In this work, we release FIGARO, a Transformer-based conditional model trained to generate symbolic music based on a sequence of high-level control codes. To this end, we propose description-to-sequence learning, which consists of automatically extracting fine-grained, human-interpretable features (the description) and training a sequence-to-sequence model to reconstruct the original sequence given only the description as input. FIGARO achieves state-of-the-art performance in multi-track symbolic music generation both in terms of style transfer and sample quality. We show that performance can be further improved by combining human-interpretable with learned features. Our extensive experimental evaluation shows that FIGARO is able to generate samples that closely adhere to the content of the input descriptions, even when they deviate significantly from the training distribution

2.3 Diffusion-LM Improves Controllable Text Generation

- `https://arxiv.org/pdf/2205.14217.pdf`
- Controlling the behavior of language models (LMs) without re-training is a major open problem in natural language generation. While recent works have demon- strated successes on controlling simple sentence attributes (e.g., sentiment), there has been little progress on complex, fine-grained controls (e.g., syntactic structure). To address this challenge, we develop a new non-autoregressive language model based on continuous diffusions that we call Diffusion-LM. Building upon the recent successes of diffusion models in continuous domains, Diffusion-LM iteratively denoises a sequence of Gaussian vectors into word vectors, yielding a sequence of intermediate latent variables. The continuous, hierarchical nature of these inter- mediate variables enables a simple gradient-based algorithm to perform complex, controllable generation tasks. We demonstrate successful control of Diffusion-LM for six challenging fine-grained control tasks, significantly outperforming prior work.

2.4 Symbolic music generation conditioned on continuous-valued emotions

- `https://arxiv.org/pdf/2203.16165.pdf`
- In this paper we present a new approach for the generation of multi-instrument symbolic music driven by musical emotion. The principal novelty of our approach centres on conditioning a state- of-the-art transformer based on continuous-valued valence and arousal labels. In addition, we provide a new large-scale dataset of symbolic music paired with emotion labels in terms of valence and arousal. We evaluate our approach in a quantitative manner in two ways, first by measuring its note prediction accuracy, and second via a regression task in the valence-arousal plane. Our results demonstrate that our proposed approaches outperform conditioning using control tokens which is representative of the current state of the art

2.5 Noise2Music: Text-conditioned Music Generation with Diffusion Models

- https://arxiv.org/abs/2302.03917
- https://google-research.github.io/noise2music/
- We introduce Noise2Music, where a series of diffusion models is trained to generate high-quality 30-second music clips from text prompts. Two types of diffusion models, a generator model, which generates an intermediate representation conditioned on text, and a cascader model, which generates high-fidelity audio conditioned on the intermediate representation and possibly the text, are trained and utilized in succession to generate high-fidelity music. We explore two options for the intermediate representation, one using a spectrogram and the other using audio with lower fidelity. We find that the generated audio is not only able to faithfully reflect key elements of the text prompt such as genre, tempo, instruments, mood, and era, but goes beyond to ground fine-grained semantics of the prompt. Pretrained large language models play a key role in this story – they are used to generate paired text for the audio of the training set and to extract embeddings of the text prompts ingested by the diffusion models.

2.6 DANCE2MIDI: DANCE-DRIVEN MULTI-INSTRUMENTS MUSIC GENERATION

- https://www.catalyzex.com/paper/arxiv:2301.09080
- Dance-driven music generation aims to generate musical pieces conditioned on dance videos. Previous works focus on monophonic or raw audio generation, while the multi- instruments scenario is under-explored. The challenges of the dance-driven multi-instruments music (MIDI) genera- tion are two-fold: 1) no publicly available multi-instruments MIDI and video paired dataset and 2) the weak correla- tion between music and video. To tackle these challenges, we build the first multi-instruments MIDI and dance paired dataset (D2MIDI). Based on our proposed dataset, we in- troduce a multi-instruments MIDI generation framework (Dance2MIDI) conditioned on dance video. Specifically, 1) to model the correlation between music and dance, we encode the dance motion using the GCN, and 2) to generate harmo- nious and coherent music, we employ Transformer to decode the MIDI sequence. We evaluate the generated music of

our framework trained on D2MIDI dataset and demonstrate that our method outperforms existing methods. The data and code are available on https://github.com/Dance2MIDI/Dance2MIDI

2.7 Moûsai: Text-to-Music Generation with Long-Context Latent Diffusion

- https://arxiv.org/pdf/2301.11757.pdf

2.8 review of music generation

- https://www.catalyzex.com/paper/arxiv:2211.09124

2.9 SYMBOLIC MUSIC GENERATION WITH DIFFUSION MODELS

- https://arxiv.org/pdf/2103.16091.pdf

2.10 Diffusion-LM on Symbolic Music Generation with Controllability (stanford)

- http://cs230.stanford.edu/projects_fall_2022/reports/16.pdf

3 Available datasets

3.1 giant-piano midi dataset

- GiantMIDI-Piano: A large-scale MIDI Dataset for Classical Piano Music

- https://arxiv.org/pdf/2010.07061.pdf

3.2 mono midi transposition dataset

- simpler dataset https://sebasgverde.github.io/mono-midi-transposition-dataset/

4 diverse

4.1 overview of different music gen methods

- https://www.catalyzex.com/s/music%20generation