Music Is all You Need for Art (Proposal)

Luke Bidulka

November 16, 2022

1 Introduction

Generative AI models have seen vast progress in recent years, notably in qualitative domains like language and image generation [7, 9, 10, 11]. The current state-of-the-art results in these areas represent a stunning increase in quality from just a few years ago [2]. The related music generation domain has seen increased interest, but has not yet experienced the success of other generative areas [8].

However, there are vanishingly few works on image generation from music, despite the fact that visual and musical experience is ubiquitously combined in mediums like music videos. Building on the successes in related areas, this is a natural next-step for AI content generation. To further this aim, we propose a novel method to generate genre and mood informed music album covers from raw audio input.

We first build a genre and mood classifier off of the VGGish architecture [5] which takes in spectrogram images of 10s input audio snippets. The genre classification is used to retrieve a genre-representing starting image and fill in a template prompt with the genre and mood strings, then perform image-to-image transfer using Stable Diffusion [11]. Thus, an output image is produced which represents the genre and mood of the input audio snippet. This pipeline is described in more detail in section 4.

2 Related Work

The pursuit of generative AI models has a long history, punctuated by recent successes which have propelled it to mainstream recognition. Large scale language models such as GPT-2 [7] and GPT-3 [9] are now capable of producing human-like text and recent models such as unCLIP (DALLE-2) [10] and Stable Diffusion [11] (SD) can create incredibly impressive images from text inputs. However, music generation has not yet seen the same success. Partly due to more strict copyright and less available data compared to image and text domains, which have truly massive datasets to work with.

There are a few large scale music datasets which provide genre labels, such as the Free Music Archive (FMA) [3] and the Million Song Dataset [1] but while the FMA contains genre labels, and the Million Song Dataset has two fields which are similar to genre and mood ("energy" and "danceability") neither are suitable for our task. There are almost no datasets which include music mood, due to its subjectivity. The exception is AudioSet [4] a dataset constructed from annotated 10-second YouTube sound clips, this is the dataset we use in our work and which is explained in detail in 3.

There has been much work on audio processing and large scale classification, such as the VGGish model, which we utilize in our work [5]. However, very few works directly tackle the problem of generative images from music data. One recent attempt was by Qie et al [6], who generated natural scene images from a small dataset of music found by natural keywords "sky", "water", "mountain", and "desert". A more advanced and recent work by Zhao et al [12] proposes an audio-to-image generative adversarial network to generate visual images directly from audio

spectrograms. However, this work focuses on filling in known bi-modal data (there are known ground-truth images for the audio), not creating a more creative representation of the audio data.

In contrast, we aim to produce visually appealing representations of audio data directly.

3 Dataset

For this work we will use Google AudioSet: a large scale collection of human annotated 10-second sound clips drawn from YouTube videos containing 2 million samples with 1 or more of 527 labels. Some example labels with the thumbnails of the corresponding YouTube videos are shown in 1. We will use the 128-dim feature version of the dataset, consisting of labeled 128-dim features vectors. These 128-dim feature vectors are the output of the VGGish acoustic model which takes spectrogram images as input. The audio preprocessing code to produce appropriate spectrogram images and the VGGish model are open sourced and can be used to process arbitrary audio input.

The dataset comes with 3 splits: eval, balanced train, unbalanced train. Eval contains 20,383 segments from distinct videos, with at least 59 examples for each of the 527 classes, though many have more than this due to label co-occurrence. Balanced train contains 22,176 segments from distinct videos chosen with at least 59 examples for each of the 527 classes in the same manner. Unbalanced train contains the remaining 2,042,985 segments from distinct videos.

Of the entire AudioSet, we will only use utilize those which are labeled with the 32 relevant music genre and mood labels. We will use the following music genre labels (25 classes, 2k-40k samples per): Pop, Hip hop, Rock, R&B, Reggae, Soul, Country, Funk, Folk, Middle Eastern, Jazz, Disco, Classical, Electronic, Latin, Blues, Children's, New-age, Vocal, Music of Africa, Christian, Music of Asia, Ska, Traditional, Independent. We will use the following music mood labels (7 classes, 1k-6k samples per): Happy, Funny, Sad, Tender, Exciting, Angry, Scary.

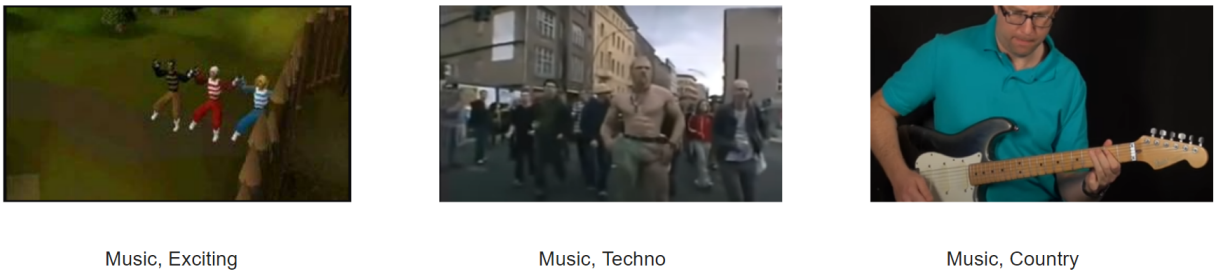

Figure 1: Some Dataset Examples, with Corresponding YouTube Thumbnails

4 Model & Methods

We propose a three step process: (1) audio feature extraction via the VGGish architecture, (2) genre and mood classification of the audio feature vector, and (3) image-to-image transfer of a genre-representing starting image based on a text prompt filled in with the proposed genre and mood labels via Stable Diffusion [11]. The total process is shown visually in 2 and explained in more detail below.

Since we are using the 128-dim feature version of the AudioSet data, we will use a simple 1-layer MLP to classify the mood and genre of the given feature vector and use the fixed VGGish architecture as the audio feature extractor, though in this project it will not be used unless time allows for additional testing on unseen raw audio inputs. The VGGish architecture is essentially the same as the standard VGG model (configuration A, with 11 weight layers) with modifications made to fit the input spectrogram data and to output a 128-dim feature vector. The architecture is further outlined in [5].

After classification of the feature vector, the output genre label is used to select a corresponding base image with

which to perform image-to-image transfer. 25 genre representative images will be available, one for each genre label. The genre starting image will be passed on to the Stable Diffusion model and paired with a text prompt to perform the transfer.

Similar to the genre representative images, 25 template text prompts will be hand created in advance to be drawn from according to the classified genre. An example prompt suitable for rock music is "lead singer of rock band who is MOOD", where MOOD will be replaced with the predicted mood label (such as "exciting") to create a full text prompt to be subsequently passed on to the Stable Diffusion model.

Image-to-image transfer is then performed using the selected genre representing image and genre prompt template filled in with mood as input to Stable Diffusion.

Thus, an output image is produced which represents the genre and mood of the input audio snippet and our process is complete.

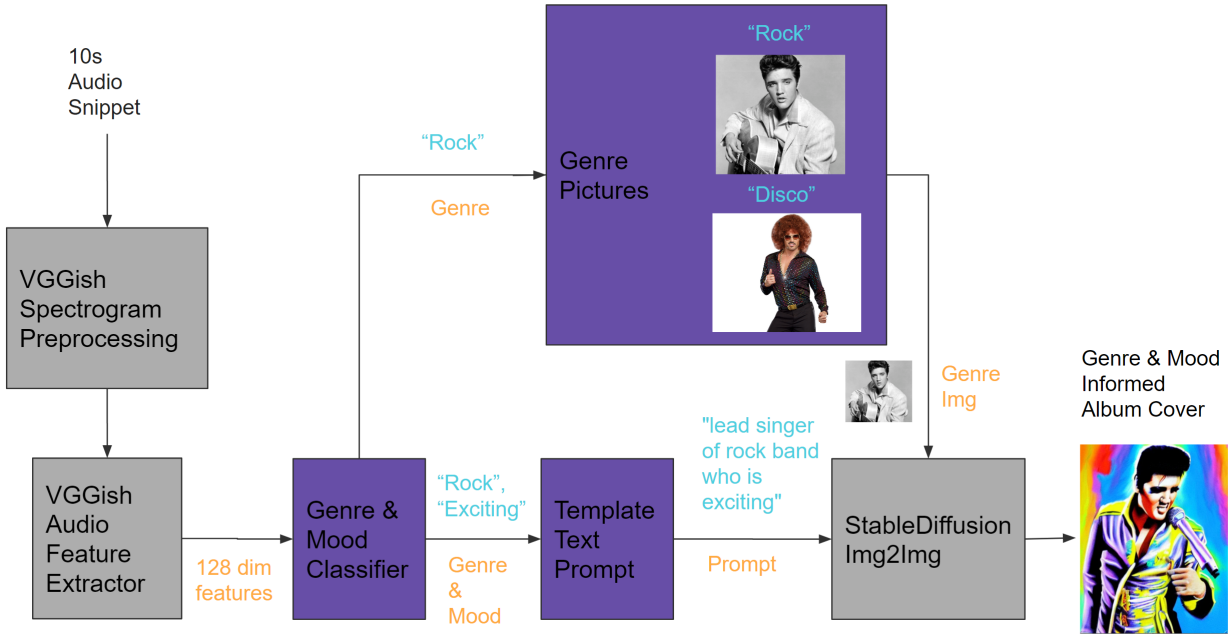

Figure 2: Proposed Pipeline

Project Timeline:

1. (Complete) Dataset initial setup, basic dataloading, initial literature review
2. (Nov 15) Proper PyTorch Dataset dataloader setup, preprocessing code
3. (Nov 18) Genre and mood classifier setup and working, no hyperparameter tuning
4. (Nov 22) Initial SD prompt structure and settings determined
5. (Nov 29) Classifier tuned, prompt determined, full pipeline done, project results generated
6. (Dec 6) Report Submission

5 Evaluation

Evaluation of our process will be done separately for the classification and the image generation components.

Classification: As in previous work such as [5] we will use F1 score and balanced average AUC over classes (probability of correct positive classification, as function of negative classification, perfect = 1.0, random = 0.5) to evaluate the classification network performance.

Image Generation: Because the quality of the image outputs is inherently qualitative/subjective, the prompt structure and SD model parameters will be selected to generate "good-looking" images according to our intuitive perception.

References

[1] Thierry Bertin-Mahieux et al. "The million song dataset". In: (2011).

[2] Karol Gregor et al. "Draw: A recurrent neural network for image generation". In: *International conference on machine learning*. PMLR. 2015, pp. 1462–1471.

[3] Michaël Defferrard et al. "FMA: A dataset for music analysis". In: *arXiv preprint arXiv:1612.01840* (2016).

[4] Jort F Gemmeke et al. "Audio set: An ontology and human-labeled dataset for audio events". In: *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE. 2017, pp. 776–780.

[5] Shawn Hershey et al. "CNN architectures for large-scale audio classification". In: *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE. 2017, pp. 131–135.

[6] Yue Qiu and Hirokatsu Kataoka. "Image generation associated with music data". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2018, pp. 2510–2513.

[7] Alec Radford et al. "Language models are unsupervised multitask learners". In: *OpenAI blog* 1.8 (2019), p. 9.

[8] Jean-Pierre Briot and François Pachet. "Deep learning for music generation: challenges and directions". In: *Neural Computing and Applications* 32.4 (2020), pp. 981–993.

[9] Tom Brown et al. "Language models are few-shot learners". In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[10] Aditya Ramesh et al. "Hierarchical text-conditional image generation with clip latents". In: *arXiv preprint arXiv:2204.06125* (2022).

[11] Robin Rombach et al. "High-resolution image synthesis with latent diffusion models". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 10684–10695.

[12] Pengcheng Zhao et al. "Generating images from audio under semantic consistency". In: *Neurocomputing* 490 (2022), pp. 93–103.