Music Is all You Need for Art (Proposal)

Luke Bidulka

December 21, 2022

Abstract

In this work, we present a novel method to create visually appealing, mood and genre informed image representations directly from music audio data. We show the use of the VGGish architecture to extract feature vector representations of input 10s YouTube video audio snippets and classify their mood and genre with two MLP networks. Evaluating the classifiers on the AudioSet dataset, we find them to perform well. We then show how the predicted mood and genre labels can be used to construct a template prompt string for input to the StableDiffusion text-to-image model to generate mood and genre informed album covers as output. Using a fine tuned GPT-2 model to extend the basic prompts to more elaborate, expressive prompts, we produce a qualitatively distinct style of output image compared to the initial template prompt. By qualitatively evaluating a number of the output album covers in comparison to their input YouTube video segments, we find the overall method to perform well.

1 Introduction

Generative AI models have seen vast progress in recent years, notably in qualitative domains like language and image generation [8, 10, 12, 13]. The current state-of-the-art results in these areas represent a stunning increase in quality from just a few years ago [3]. The related music generation domain has seen increased interest, but has not yet experienced the success of other generative areas [9].

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 use the VGGish architecture [6] to produce 128x1 condensed feature representations of input spectrogram images of 10s input nusic audio snippets. The feature vector then has its genre and mood classifier using two neural networks, and the predicted genre and mood labels are substituted into a template string to be used as an image generation prompt for StableDiffusion. The initial prompt is extended using a fine-tuned GPT-2 model to be more expressive and draw out better images, and both the initial and extended prompts are fed into StableDiffusion to produce output images. The images generated from the initial, simpler prompt are considered the "Album" style outputs, and those generated from the extended prompts are considered the "Conceptual" style outputs. 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 Sec 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 [8] and GPT-3 [10] are now capable of producing human-like text and recent models such as unCLIP (DALLE-2) [12] and Stable Diffusion [13] (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) [4] 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 [5] 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 [6]. However, very few works directly tackle the problem of generative images from music data. One recent attempt was by Qie et al [7], 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 [15] 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 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 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 [6]. 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.

Figure 1: Some Dataset Examples, with Corresponding YouTube Thumbnails

3.1 Dataset Preprocessing

Of the entire AudioSet, we only utilize those which are labeled with at least one of the 32 relevant music genre and mood labels. We 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 use the following music mood labels (7 classes, 1k-6k samples per): Happy, Funny, Sad, Tender, Exciting, Angry, Scary.

The reduced Music Mood and Music Genre labels only dataset resulted in 1,407,804 samples with Mood labels and 169,360 samples with Genre labels. Like the original dataset, the distribution of the reduced set labels was unequal as shown in Fig 2. The original AudioSet splits of balanced train, unbalanced train, and eval are maintained.

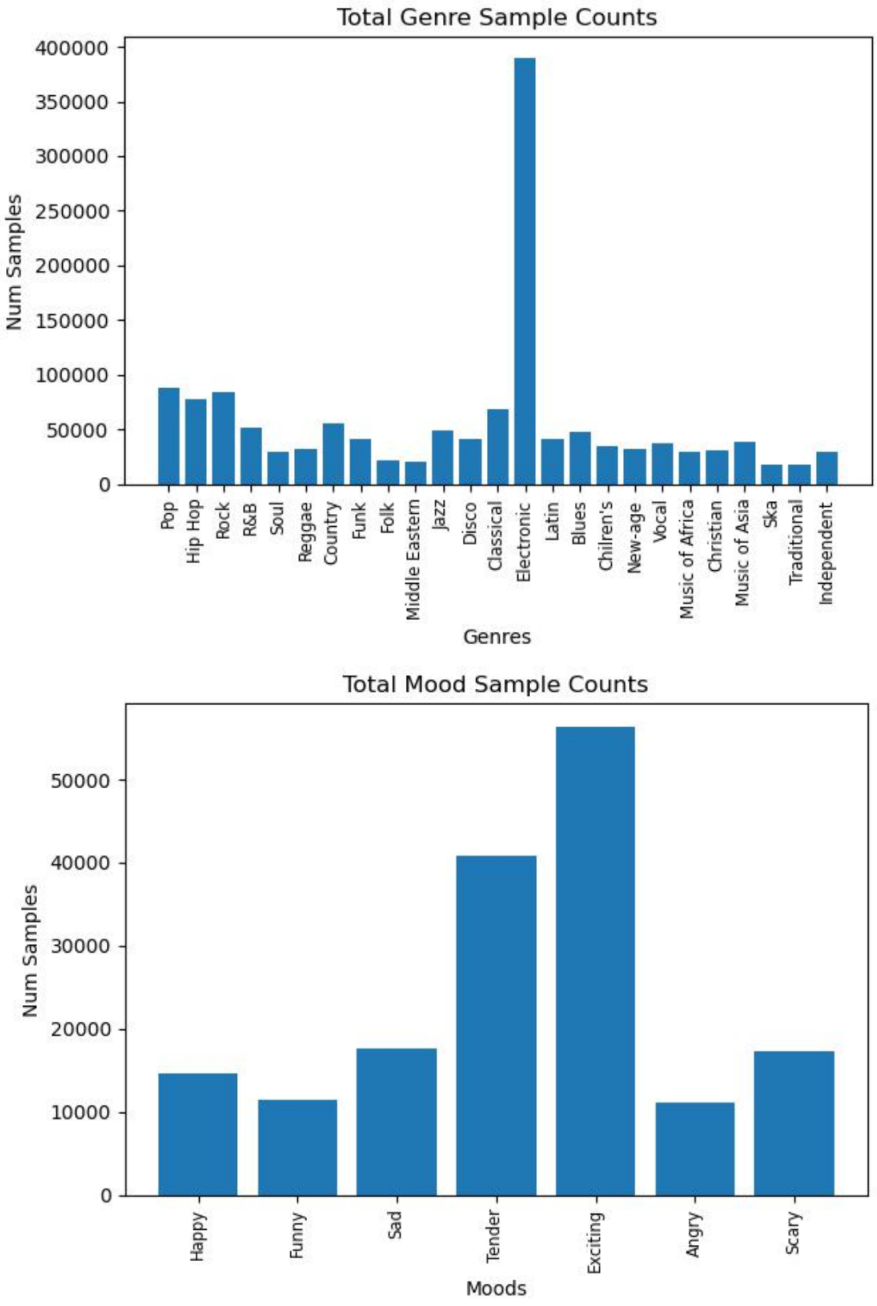

Figure 2: Total sample distribution of AudioSet data with Genre labels (top) and Mood labels (bottom).

4 Model & Methods

We propose a three step process: (1) audio feature extraction via the VGGish architecture, (2) prompt construction via genre and mood classification of the audio feature vector, and (3) text-based image generation from the constructed prompt as well as an extended prompt. The total process is shown visually in Fig 3 and explained in more detail below.

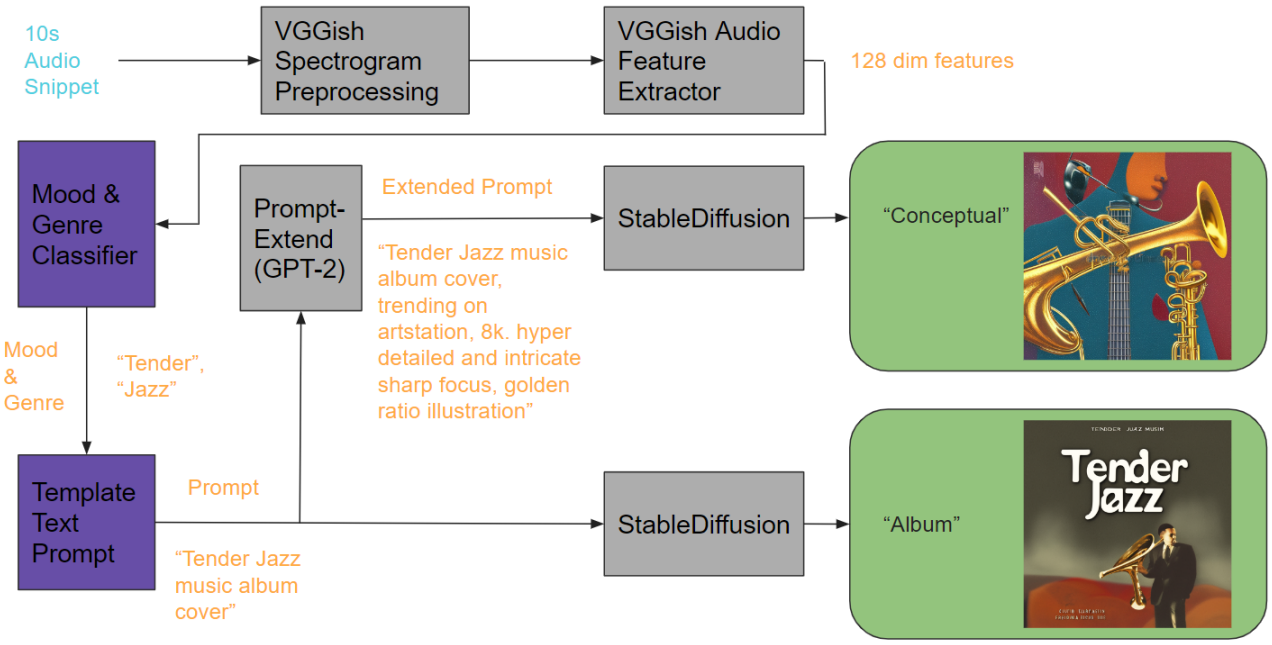

Figure 3: The overall pipeline. (1) 128-dim feature vectors are extracted from an input 10s audio snippet using VGGish, (2) the Mood and Genre of the vectors are classified, and the template prompt string of the form "MOOD GENRE music album cover" is constructed with the predicted labels, (3) the prompt is extended into a more extensive prompt using the Prompt-Extend model, the basic and extended prompts are passed through Stable-Diffusion 2.0 to create the "Album" and "Conceptual" output images respectively.

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 dimensions and to output a 128-dim feature vector. In our work, VGGish is used as a frozen feature extractor and is not modified from the architecture introduced in [6]. As outlined in Sec 3, VGGish was used to produce the AudioSet 128-dim feature vector data with which we worked.

Parts (2) and (3) of our proposed process are now explained in Sec 4.1 and Sec 4.2 respectively.

4.1 Classifiers & Template Label Construction

To construct a prompt to be fed into the generative network in step (3), we first classify the Mood and Genre of the input feature vector. With the predicted Mood and Genre labels, we fill in the template prompt which has the form: "MOOD GENRE music album cover" (eg: "Tender Jazz music album cover").

To predict the Mood and Genre labels, we construct a two classification networks: one for Mood and one for Genre. Each network is a simple MLP with one hidden layer and a batchNorm layer both of size 1024. This architecture is shown in Fig 4.

After classification of the feature vector and creation of the template prompt string, the prompt string is passed to step (3): the creation of the "Album" output image via the template prompt, and the extension of the template prompt and subsequent creation of the "Conceptual" output image.

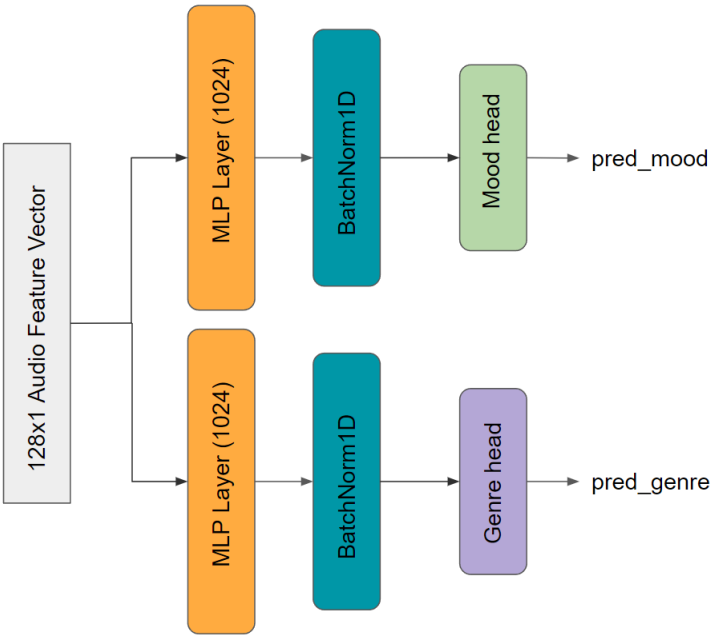

Figure 4: The Mood and Genre classification networks.

4.2 Prompt Extension & Image Generation

To produce our output album covers, we use Stable Diffusion (SD) 2.0 [13] as a text to image generator. Two outputs are created: the "Album" output which is the SD output image generated from the basic template prompt constructed in step (2) (Sec 4.1), and the "Conceptual" output which is the output image generated from an extended version of the basic template prompt.

The extended prompt is created using the Prompt Extend fine-tuned GPT-2 model from [11]. It was fine-tuned on a subset (unique prompts only) of the DiffusionDB dataset [14], a publicly available large-scale text-to-image prompt dataset which contains 14 million images generated by Stable Diffusion using prompts and hyperparameters specified by users. The model simply takes an input string and extends it in the style of an SD prompt. For example, "Tender Jazz music album cover" might be extended to "Tender Jazz music album cover, trending on artstation, 8k, hyper detailed and intricate sharp focus, golden ratio illustration". Once the basic and extended text prompts are created, they are each fed into SD and produce the Album and Conceptual output images respectively. The Prompt Extend model is used in our work as-is, for inference only.

Figure 5: Album style Stable Diffusion output images created from the template prompt "MOOD GENRE music album cover" with "Sad Electronic" (left), "Sad New-age" (middle left), "Angry Rock" (middle right), and "Exciting Children's" (right).

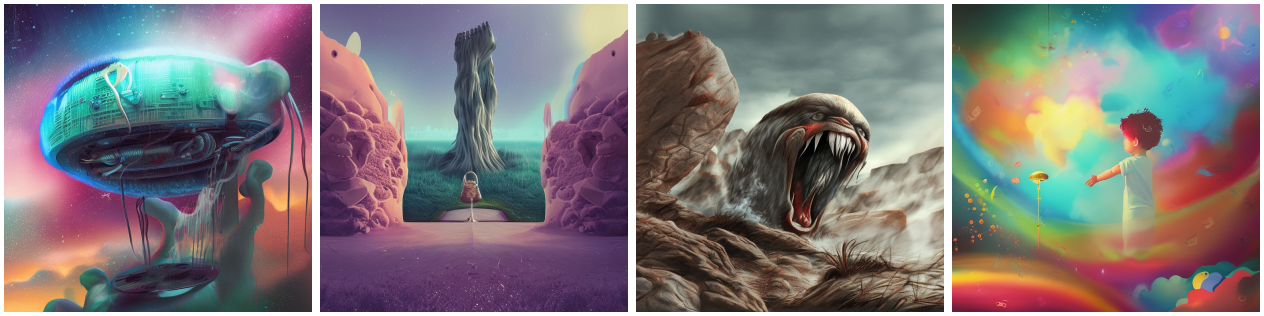

Figure 6: Conceptual style Stable Diffusion output images created from the extended template prompt "MOOD GENRE music album cover" with "Sad Electronic" (left), "Sad New-age" (middle left), "Angry Rock" (middle right), and "Exciting Children's" (right).

5 Experiments

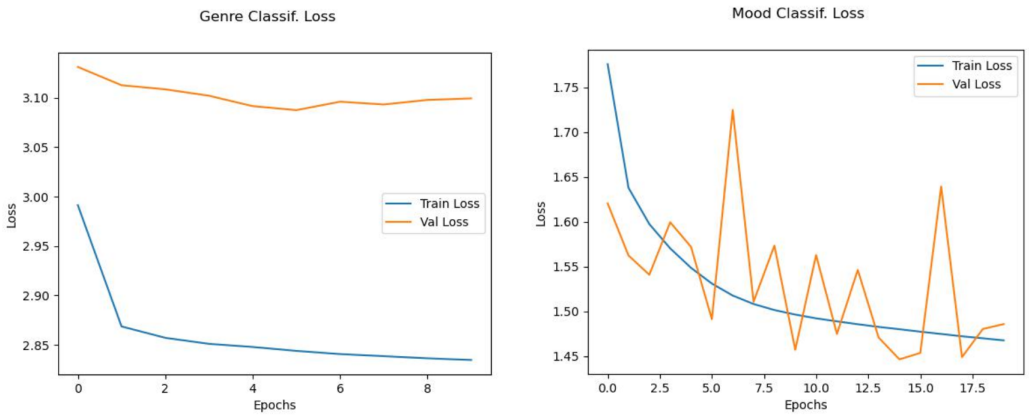

Figure 7: Training loss curves for Genre (left) and Mood (right) classification.

To evaluate our proposed method, we use the unbalanced training split of the preprocessed dataset to train the Mood and Genre classifiers (Sec 5.1). We report the F1 scores for each of the classes in the Mood and Genre tasks in Sec 5.2 and we discuss the qualitative evaluation of the SD output images for each of the two styles of prompts in Sec 5.3.

5.1 Classifier Training

The Mood and Genre classifiers were trained using the unbalanced training data split in a multi-label training setup with CrossEntropy loss and the ADAM optimizer [2] (lr = 3e-4, batch size = 4096). The resulting training loss curves are shown in Fig 7. Genre began to overfit around 10 epochs, and Mood began to overfit around 20 epochs, so training was halted at those limits. Training was done on a single Nvidia 2060 GPU.

5.2 Classifier Evaluation

Evaluation of our method is done quantitatively with per-class F1 scores for the classifiers, and qualitatively for the output image generation.

As in previous work such as [6] we use F1 score as an appropriate scoring method on unbalanced data:

$$F1 = 2 * \frac{precision * recall}{precision + recall}$$

where

$$precision = \frac{true\_positives}{true\_positives + false\_positives}$$

$$recall = \frac{true\_positives}{true\_positives + false\_negatives}$$

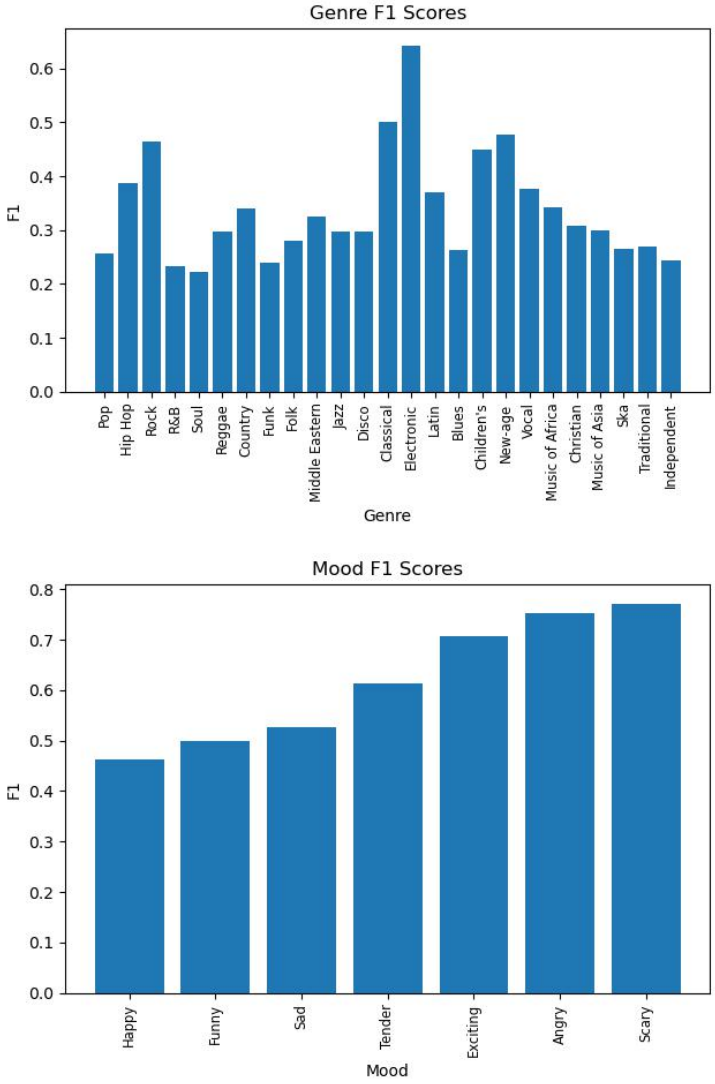

Figure 8: F1 scores per-class for Genre (top) and Mood (bottom) classification.

The eval set F1 scores for Mood and Genre for each class are shown Fig 8. We can see that the Mood classifier, as hinted by its training curve, achieves respectable results across all of the Mood classes. The Genre classifier is more strongly affected by the dataset imbalance, but still achieves respectable results.

5.3 Image Generation Evaluation

Because the objective of our work (the creation of good-looking album covers) is subjective, the evaluation of the quality of the image outputs is inherently qualitative. To evaluate the quality, we examine some examples of both

"Album" style output images (shown in Fig 5), and "Conceptual" style (shown in Fig 6). Many more examples of both styles, with acompanying prompts are shown in Appendix **??** [ADD APPENDIX WITH MORE RESULTS AND PROMPTS].

Figure 9: All Stable Diffusion output images created for this input YouTube video from the template prompt: "Angry Pop music album cover" (left and middle left), and extended prompt: "Angry Pop music album cover, 1970s sci-fi art by Neil Armstrong and James Jean, asymmetrical, Organic Painting, Matte" (middle right and right).

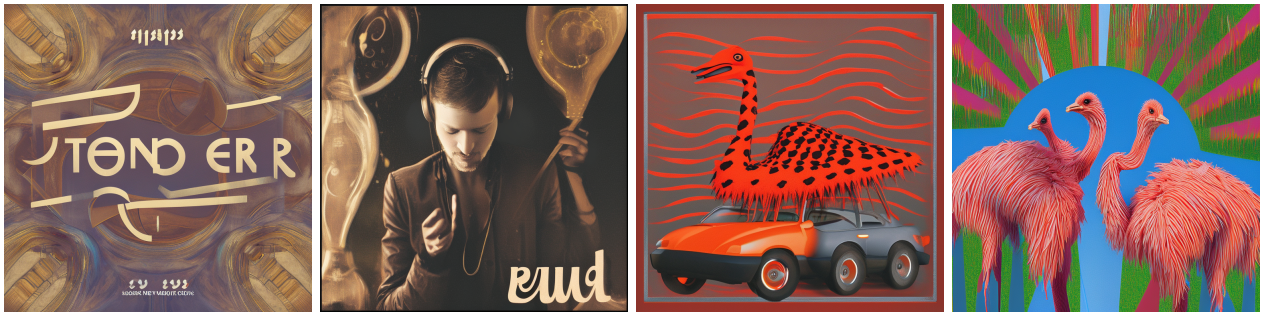

Figure 10: All Stable Diffusion output images created for this input YouTube video from the template prompt: "Tender New-age music album cover" (left and middle left), and extended prompt: "Tender New-age music album cover, the band is a red striped ostrich, the car is orange, 3D," (middle right and right).

The Album style outputs display more stylized, minimalist, and 2D properties, in contrast to the Conceptual style outputs which are more detailed and generally portray a "scene". Not every SD output image has the same qualitative appeal, so 2 Album and 2 Conceptual style images were produced per prompt input to SD. More than 2 at a time was not possible due to RAM limitations in the experimental setup, but outputting a greater number of images per input prompt would be best in practice. It would allow for selection of the most desirable outputs, and would provide a greater variety of interpretations to choose from per run.

To evaluate the goal of representing the original input audio, we consider how the output images of two typical samples represent the input YouTube video segment. The first YouTube clip is from 10s to 20s and is labeled with mood: "Angry" and genre: "Pop". Its two Album and two concept output images are shown in Fig 9. The second YouTube clip is from 240s to 250s and is labeled with mood: "Tender" and genre: "New-age". Its two Album and two concept output images are shown in Fig 10. Qualitatively, the authors find the images to represent the audio well, especially the first Album style and first Concept style images. All the images are visually appealing, with good colour contrast and striking style. Note the variety of extended prompt, and how sometimes elaborates specific ideas such as the "red striped ostrich" of Fig 10.

Overall, the quality of generated images is very high and thus we subjectively consider our objective to be well satisfied.

6 Conclusion

In this work, we present possibly the first method to create visually appealing representations directly from music audio data. We used Google's AudioSet dataset to demonstrate the Mood and Genre classification of VGGish extracted feature vectors and their use in constructing input prompts for the Stable Diffusion text-to-image generator. Quantitatively evaluating the F1 scores of the classification networks showed acceptable performance across all classes, but revealed the negative impact of the highly imbalanced training data. By subjectively evaluating a range of generated output images, their aesthetic appeal was found to achieve our overall goal of creating good looking album covers directly from music audio.

There are clear and immediate next steps to improve this work, most of which stem from the limited time allocated for this project. First, the imbalanced training data should be addressed using methods such as weighting the loss by number of samples per class or by enforcing the equal frequency of samples from each class during training. Accounting for the imbalanced training data will certainly improve the classification performances, especially for the Genre classification which suffered the most from this problem. Further, a more extensive subjective analysis of the output images should be performed, to find out how often the output images are found to represent the input audio sequence. Due to the limited time of this project, this analysis was limited to a few hand-picked samples.

Longer term, this work could be extended to have a more advanced generative stage. Using image in-painting for example, arbitrary text (eg: Album title and artist name) could be placed onto the output images. An additional visual processing stream could capture the visual content in the input videos and be used to further condition the image generation prompts.

References

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

[2] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

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

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

[5] 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.

[6] 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.

[7] 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.

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

[9] 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.

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

[11] Das. *Prompt Extend*. Accessed Dec 19, 2022 https://github.com/daspartho/prompt-extend. 2022.

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

[13] 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.

[14] Zijie J. Wang et al. "DiffusionDB: A Large-Scale Prompt Gallery Dataset for Text-to-Image Generative Models". In: *arXiv:2210.14896 [cs]* (2022). URL: https://arxiv.org/abs/2210.14896.

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