A StyleGAN-2 inspired Generative Adversarial Network for the PCA-controllable generation of drums samples for content-based retrieval

Alejandro Koretzky
Splice
ale@splice.com

Naveen Rajashekharappa
Splice
naveen@splice.com

ABSTRACT

Generative Adversarial Networks (GAN) have proven incredibly effective at the task of generating highly realistic natural images. On top of this, approaches for the conditioning of the generation process by controlling specific attributes in the latent space (e.g. hair color, gender, age, beard, etc when trained on human faces) have been gaining more attention in recent years. In this work, we validate a StyleGAN-2 inspired architecture for the unlimited generation of high-quality magnitude spectrogram images, for the purpose of content-based retrieval. In addition, in the same way that it is possible to discover and control specific attributes relevant to the distribution of natural images, we demonstrate that the same is applicable to the domain of audio, showing that when trained on drum loops, some of these controllable latent dimensions directly relate to highly semantic factors such as BPM, rhythmic pattern, low pass and high pass filtering, etc. Even though these generated high-resolution spectrograms can be inverted back into the time-domain and made available for use (we demonstrate this using the Griffin-Lim algorithm), the purpose of this project was to validate the approach with the goal of content-based retrieval. Particularly, developing better search and discovery tools for querying a large collection of human-made audio samples.

1. DESIGN

1.1 Generative model

Using NVIDIA's StyleGAN-2 model architecture as a baseline, we modify the model to accommodate rectangular input shapes of 256 x 512 pixels, where 256 corresponds to a 512-point DFT used in the computation of the magnitude Short Time Fourier Transform (STFT). We train this modified model on 45.000 examples of complex drum loop samples drawn from a diverse set of musical genres. Upon convergence, our model has learned to generate highly-detailed magnitude spectrograms of drums content, indistinguishable from real ones.

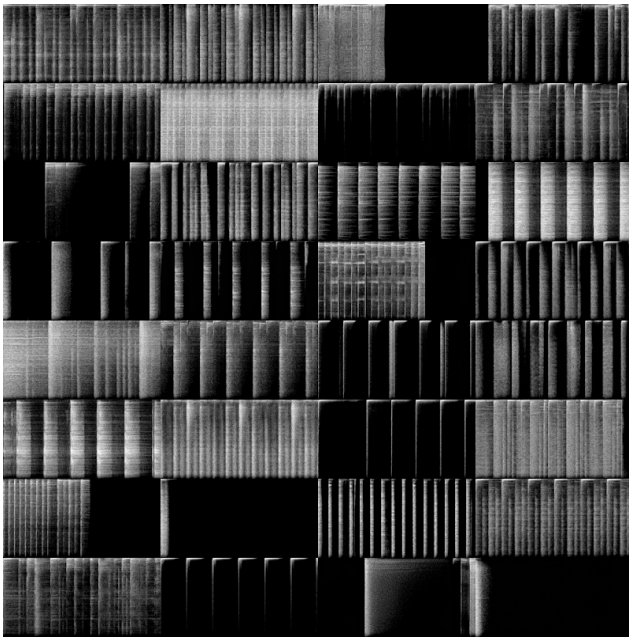

Figure 1. Multiple sample outputs from our generative drums model.

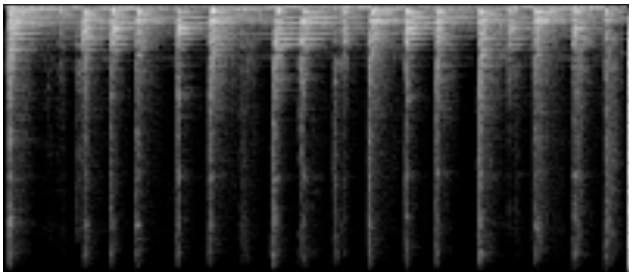

Figure 2. Single 256 x 512 output from our generative drums model.

2. CONDITIONAL GENERATION

One of the open challenges with GAN models is being able to obtain highly semantic and disentangled latent space dimensions, so that we can control the generative process with a clear and intuitive understanding of the attributes being modified. While several approaches have been pro-

posed, we validated the use of PCA projections of the latent space representation in order to manipulate the generative process in a semantic way. We found that, just like in the case of natural images, the PCA bases happen to control highly semantic attributes related to the specific domain, in this case, drum loop sounds. For instance, the first few bases control the overall structure, rhythmic pattern and BPM of the generated drum loops. As we start moving to higher bases, we start controlling more subtle, fine-grained attributes of the generated samples such as small structure variations, the addition of specific elements (for example, one basis controlled the addition of hi-hat sounds to the generated pattern), and even low pass and high-pass filtering on the generated outputs. While some of the bases did not fully capture a clear attribute or behavior (some bases would drastically change the output past a certain range), the approach validates some key parallels with the domain of natural images.

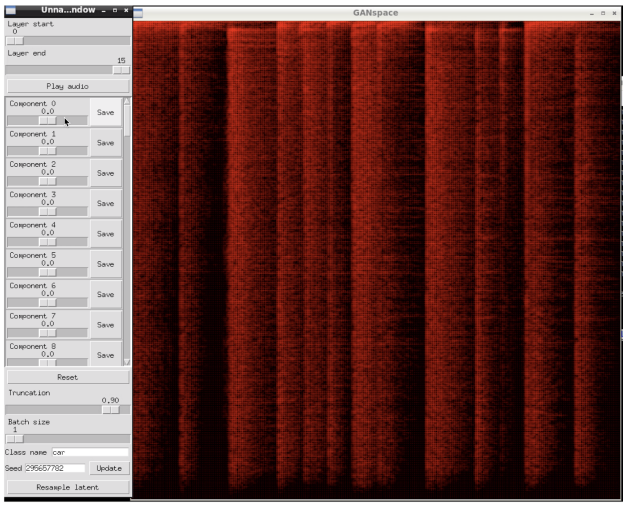

Figure 3. Interface for manipulating the generative latent space using PCA projections. Each PCA basis can be controlled using a slider. As the values change, the generated magnitude spectrogram (red) starts changing based on the underlying attributes being controlled by the corresponding basis. (video available as part of the submission)

3. CONTENT-BASED RETRIEVAL

Using the generated outputs from our StyleGAN-2 inspired model for generating drum loops, we perform similarity search against a large collection of human-made drum loops from the Splice platform catalog. The results clearly show that a synthetically generated sample can effectively be used for content-based retrieval against a collection of real, human made samples. In addition, this important validation illustrates an interesting use case in this new era of AI-generated content, where rather than just relying on generative models for fulfilling our content needs, we combine generative and content-based retrieval capabilities to streamline the discovery of human-made content beyond the ubiquitous tag-based search.

4. REFERENCES

[1] T. Karras, S. Laine, T. Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. *CVPR*. 2019

[2] E. Härkönen, A. Hertzmann, J. Lehtinen, S. Paris. GANSpace: Discovering Interpretable GAN Controls. NeurIPS. 2020

[3] R. Abdal, P. Zhu, N. Mitra, and P. Wonka. Styleflow: Attribute-conditioned exploration of stylegan-generated images using conditional continuous normalizing flows. arXiV:2008.02401, 2020.

[4] T. Karras, S. Laine, M. Aittala, J. Hellsten and J. Lehtinen, T. Aila. Analyzing and Improving the Image Quality of StyleGAN. *Proc. CVPR*. 2020