INSTRUMENTAL AUDIO SYNTHESIS USING GANS PROJECT PROPOSAL

Group N
Etienne Leclerc (V00853992), Jordie Shier (V00688891), Lu Lu (V00836042),
Yangruirui Wang (V00949204), and Ziyi Feng (V00940985)

1 PROBLEM DEFINITION

Music producers and sound-effect artists require the frequent use of audio samples, for example to simulate a snare drum hit or a gunshot sound. The process of arduously searching through multi-gigabyte sound libraries for the right sound is extremely time-consuming; and its alternative, of creating a brand new sound using audio synthesizers, requires a formidable technical background. To compound this difficulty, in a track or action scene containing hundreds of snare drum hits or gunshots, the process of using the same (or same few) samples repeatedly can lead to a canned effect. What is required is a method for simultaneously trimming down the size of a library while simultaneously (and in some sense paradoxically) increasing sound variability.

The problem we would like to tackle in our project is using Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) to generate new instrumental audio from a dataset of existing material. GANs have the potential to be used to generate new sounds on the fly, and can be used to interpolate between sounds to help music producers hone in on a desired sound. This would dramatically alleviate both the problem of having to pore through giant sound libraries, and the problem with having to only use one sample repeatedly. In addition, the explosion of new sounds which could potentially be produced by GANs would vastly reduce recording costs by designers of sound libraries.

This research avenue is to a certain degree untapped: GANs have been successfully applied to the generation and manipulation of images, however, relatively little work has been focused on the audio domain. Research related to the specific work proposed here was presented by Donahue et al. (2018) and Engel et al. (2018).

The goal of our project will be to implement a GAN, similar to the one proposed by Donahue et al. (2018), and train it using audio samples recorded from instrumental sounds including brass, string, reed, and mallet instruments. We will use the freely available NSynth dataset[1], published by Engel et al. (2017) from Google's Magenta research lab. The NSynth dataset contains over 300k four second long audio samples of labelled musical instruments.

2 GOALS

As described in the first section, our goal in this project is to use Generative Adversarial Networks (GANs) to generate new instrumental audio from the NSynth dataset. There are two main components in a GAN: the generator network that is tasked with generating new material, and a discriminator network that has been pre-trained on a dataset and is tasked with classifying input data as real or fake. During training the generator learns how to fool the discriminator and create realistic material. In our project we will train the discriminator on short audio samples of solo instrumental audio with the goal of producing a generator that can create new instrumental sounds.

From our research so far, we have learned GANs are challenging to evaluate using objective measurements, although some metrics have been proposed. Donahue et al. (2018) used two objective measures: *inception score* and *nearest neighbour* comparisons and we will use both of those measurements.

[1] `https://magenta.tensorflow.org/datasets/nsynth`

In addition to those objective measurements, Donahue et al. (2018) carried out a subjective evaluation to further verify the quality of their results. While performing a full subjective assessment is beyond the scope of this class project, we would like to create a simple 'Turing test' to share with some friends and family as an informal evaluation.

If we are ahead of schedule, we will try a variation on the generator and discriminator. Then will have two generators and two discriminators and can do a comparison on them to find out which combination is better. An additional stretch goal will be to implement an interface for interpolating between generated sounds.

3 PLAN

There are five main stages for keeping track of our progress. The first stage is to do relevant research, and this stage will include checking out certain papers or books and talking to certain people who are familiar with this topic. After we get enough information, we will implement and train the GAN. This implementation state will occur for a little over two weeks. Then the next two stages are about evaluating the model using objective and subjective methods. Stage 5 is a stretch goal if we have extra time. And lastly, we will wrap up what we have done and write the final report. Throughout our project, to ensure everyone is involved and participates, we will hold a Zoom meeting every Friday to share new information and set goals for the next week. We also are engaging in regular conversation through a Slack workspace.

3.1 STAGE 1: RESEARCH

Dates: June 12 - July 3

As GANs are a new topic for all of us, research and learning will be a large part of the early stages of our project. During this phase of our work we will be conducting independent research and sharing our findings as informal presentations during our weekly Zoom call. We have put together a list of resources that will be helpful for us throughout our project:

- Online tutorial on audio synthesis with GANs (Pasini, 2019)
- Related papers on audio synthesis using GANs (Donahue et al., 2018; Engel et al., 2018)
- Deep learning textbook (LeCun et al., 2015)
- Machine Learning Mastery (this tutorial on GAN latent space interpolation could be helpful for the stretch goal of interpolating between sounds) (Brownlee, 2019)
- Original GAN paper (Goodfellow et al., 2014)
- Tutorial on GANs (Goodfellow, 2016)
- Additionally, Jordie's supervisor George Tzanetakis will be a valuable resource for questions regarding audio processing

3.2 STAGE 2: IMPLEMENTATION

Dates: July 4 - July 20

The goal of the implementation stage is to produce a functioning GAN that can generate new instrumental audio after being trained on audio samples from the NSynth dataset. We plan on using TensorFlow[2] to implement a model similar to the one proposed by (Donahue et al., 2018). Implementation will be broken into four subsections: data preparation, generator implementation, discriminator implementation, and integration. Data preparation will consist of selecting the instruments that we want to use to train the GAN on. We plan on trying several different types of selections such as training using brass sounds only, or training using a combination of different instrumental sounds.

Our team will split up into two smaller groups to implement the GAN. One team will implement the generator network and another team will implement and train the discriminator network. Once both

[2] `https://www.tensorflow.org/`

of these networks are constructed and the discriminator is trained, we will integrate them and train the full GAN network.

In addition to experimenting with training with different datasets, we would like to experiment with tuning network hyper-parameters and compare results. The amount that we will be able to do this will depend on how long it will take to train the network, which is an unknown factor to us at this point, although we are anticipating several hours to train each model.

Our main goal is to implement one GAN model, however, if we are ahead of schedule at this point we would like to implement a variation on the GAN proposed by (Donahue et al., 2018) by using Mel-Frequency Cepstral Coefficients, which are a type of audio transform that are more perceptually relevant to human hearing and could potentially produce improved results.

3.3 Stage 3: Objective Evaluation

Dates: July 21 - July 27

To objectively measure the quality of our implemented GAN we will use inception score and nearest neighbour evaluation, which was used by Donahue et al. (2018). Inception score can be used to evaluate whether the sound generated by the generator is close to the real sound (Xu et al., 2018). We can calculate the similarity rate or score from this in order to evaluate and optimize our generator.

3.4 Stage 4: Informal Subjective Evaluation

Dates: July 21 - July 27

Create a simple *Turing Test* to share with some friends and family as an informal evaluation. In this informal evaluation we will share the audio results generated from our GAN and get subjects to rate the quality.

3.5 Stage 5: Interpolation (Stretch Goal)

Dates: July 21 - July 27

Interpolation is a stretch goal that we would like to implement if we have time. The goal of this stage would be to provide a way to interpolate between sounds by linearly moving between two points in the generator latent space. This goal was inspired by image GANs that show smooth interpolation between different faces. Being able to smoothly adjust and move between different sounds would be a useful feature for music producers and sound effect designers. The tutorial by Brownlee (2019) will be helpful at this stage.

3.6 Stage 6: Project Report

Dates: July 28 - August 1

Wrap up project and document results in the final project report.

4 Task Breakdown

Table 1 highlights the contributions that each member of our team will make towards the completion of our project. The generator and discriminator tasks refer to implementing each of those models using the TensorFlow library. Additionally, for the discriminator task, that will also include training on selected subsets of the NSynth dataset. Preparation of the training material from NSynth is included in the data preparation task. The integration task involves putting the generator and discriminator together and training the generator. The generator will then be evaluated by all group members using two objective methods as well as an informal subjective evaluation.

Table 1: Project Task Breakdown

	Generator	**Discriminator**	**Data Preparation**	**Integration**	**Evaluation**
Etienne Leclerc		✓	✓	✓	Informal Subjective, Inception
Jordie Shier		✓	✓	✓	Informal Subjective, Inception
Lu Lu	✓		✓		Informal Subjective
Yangruirui Wang		✓	✓		Informal Subjective, Nearest Neighbour
Ziyi Feng	✓		✓		Informal Subjective, Nearest Neighbour

REFERENCES

Jason Brownlee. How to explore the gan latent space when generating faces, 2019. URL `https://machinelearningmastery.com/how-to-interpolate-and-perform-vector-arithmetic-with-faces-using-a-generative-adversarial-network/`.

Chris Donahue, Julian McAuley, and Miller Puckette. Adversarial audio synthesis. In *International Conference on Learning Representations*, 2018.

Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Douglas Eck, Karen Simonyan, and Mohammad Norouzi. Neural audio synthesis of musical notes with wavenet autoencoders, 2017.

Jesse Engel, Kumar Krishna Agrawal, Shuo Chen, Ishaan Gulrajani, Chris Donahue, and Adam Roberts. Gansynth: Adversarial neural audio synthesis. In *International Conference on Learning Representations*, 2018.

Ian Goodfellow. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Marco Pasini. Synthesizing audio with generative adversarial networks, 2019. URL `https://towardsdatascience.com/synthesizing-audio-with-generative-adversarial-networks-8e0308184edd`.

Qiantong Xu, Gao Huang, Yang Yuan, Chuan Guo, Yu Sun, Felix Wu, and Kilian Weinberger. An empirical study on evaluation metrics of generative adversarial networks. *arXiv preprint arXiv:1806.07755*, 2018.