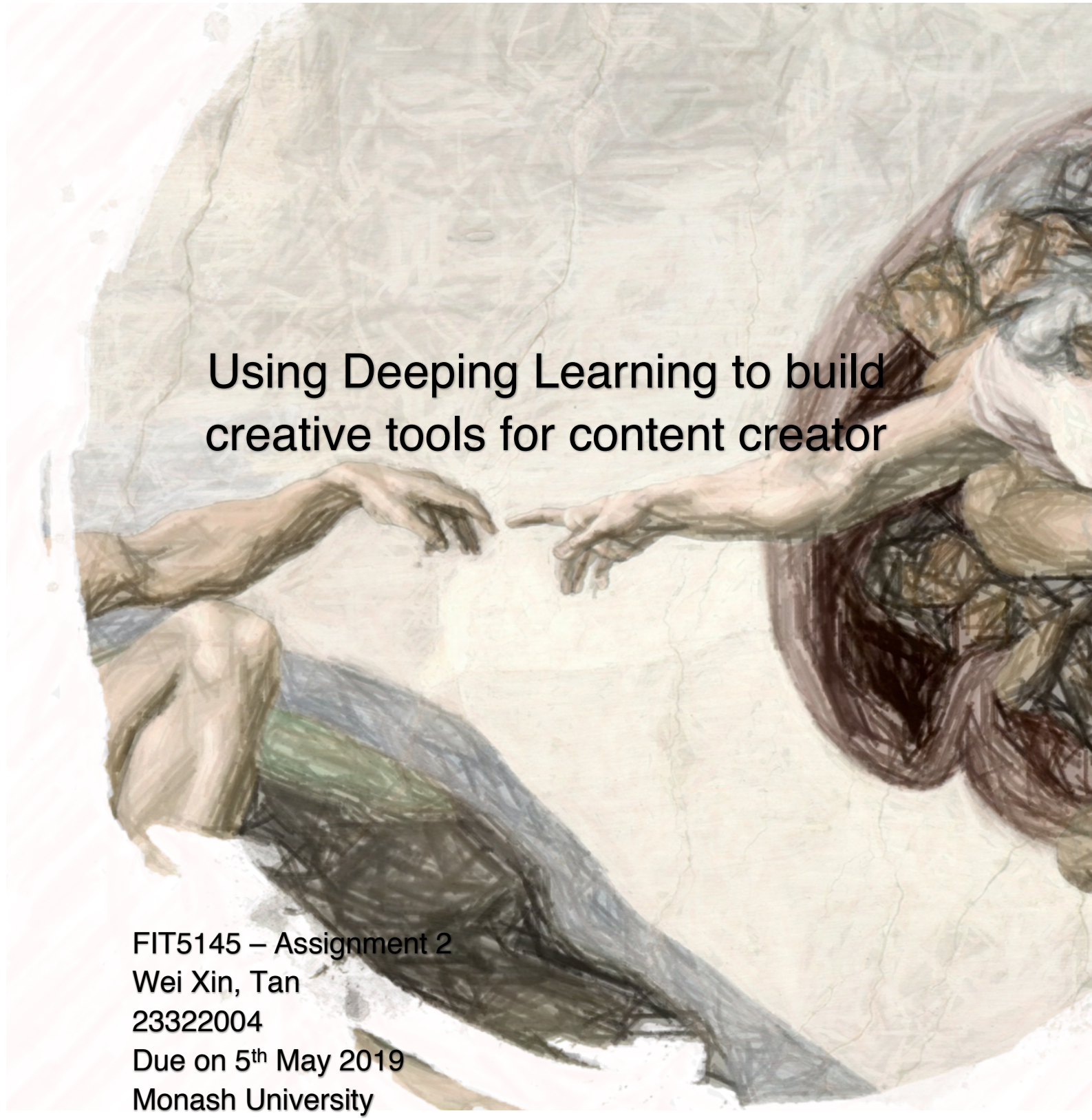

Using Deeping Learning to build creative tools for content creator

FIT5145 – Assignment 2
Wei Xin, Tan
23322004
Due on 5th May 2019
Monash University

Table of Contents

1. Project Description

1.1 Proposal

Creative work like producing art or composing music has always been deemed as work that only human can do. Because we think that machine cannot perceive emotion hence incapable of producing songs or painting that human can enjoy or relate to. Ever since a class of neural network called Generative adversarial network (GAN) was introduced by Ian Goodfellow in 2014, there are interesting development and applications in synthetic image and audio. Its applications include reconstructing 3D objects, deepfake, texture or style synthesis on image and composing instrumental music. These synthesis samples generated by GAN can be indistinguishable for the human eyes or ears.

My proposed project would be using GAN models to build creative tools for generating synthetic background image and music. These tools can be built into current existing video or photo editors. Hence, this project could be applied to Apple's Final Cut Pro, Adobe's Premiere Pro or even Youtube. Nevertheless, upon writing this project, I have never used these editing softwares because I am not a content creator or artist. Hence, I am in no position to comment on the novelty of this project. The proposal is merely my own idea, and not based on any other project but I am also not suggesting this has never been done before.

The motivation behind this project is to help small size or individual content creator to increase their production values. Firstly, due to copyright, content creators on Youtube are not allowed to use other artists' music for free. They must obtain some form of license or right, else they might need to share the profit of the video they created. Secondly, some content creator might not have the resources or capacity to create high quality thumbnail for their video. Therefore, the proposed tools will allow them to create their own background image or music. Creators who want to stand out from the crowd would also be happy for the inclusion of these tools.

1.2 Data Science Roles

The data science roles that will be involved in this project is categorized in four types and their responsibilities are discussed as follows:

Data Developer (Roles: Developer, Engineer)

Responsibilities:

1) Data Developer need to figure out the data structure and database/file system to efficiently store all the songs or images and their metadata, so that it can be easily extracted to train models on.

2) Data Developer need to consider how handle data with different format, size, et cetera. For instance, music files are stored in different format such as mp3, flac, while images have different resolutions and format.

Data researcher (Roles: Researcher)

Responsibilities:

1) Data researcher will spend their time understanding state-of-the-art models or current industrial practice, so that they can propose concept or ideas that are most viable to build the product.

Data Businessperson (Roles: Business person)

Responsibilities:

1) Data Businessperson's work will involve around convincing management or stakeholders how the proposed product could impact the company in both short and long-term. They will take into account of timing, budget, revenue, resources, market competition and so on.

Data Creative (Roles: Artist, jack-of-all-trades)

Responsibilities:

1) Since Data Creative understand wider range of tools, they will explore and work on how to deploy the models into productions like web, mobile and pc application.

However, these roles will constantly work together instead of separately, and sometimes they will even crossover their work.

2. Business Model

2.1 Actors in the Information Value Chain

Following the descriptions defined by NIST Reference Architecture (2015), the proposed project would act more as “Data consumer” than the other major actors like “Data providers”, “Big data application providers”, and “Big data framework providers” in the information value chain. Firstly, we will get all the required data (images and song tracks) from ‘Data providers’, then “Big data applications providers” would curate the data to the required format or specification in order for us to train the GAN model, hence, we are considered to be “Data consumer”. Nevertheless, these ‘providers’ mentioned earlier can be in-house or external.

2.2 Big Data Landscape

Although there are many different landscapes or ecosystems in the big data world, but our project fits into the description of ‘Application’ in most of them, and it will be in a form of “software as a service” project. Another thing to note is many data science projects used machine learning for predictive task or analysis, however, our project will be using machine learning for generative purpose, which will discuss in detail in “Data Analytic” section.

2.3 Business Model

According to the interpretation of Harvard Business Review (2012) on big-data business model, our project would more akin to “information-based differentiation” model, since it utilizes data to create differentiated service to customers. The proposed project will likely be dwelled in the “special” or “creative” project department, because it is not the core feature of the application, unless the management deem otherwise.

2.4 Potential Challenge

In foresight, I think there are three big challenges that this project will face. First is the implementation and design, specifically how UI designer, front- and back- end programmer will have to work together to deploy a product that the users will find it intuitive, and appreciate its addition. The second challenge is very common in data science project: getting enough of the required data to successfully train the model. The last challenge would be the modelling stage, as we know in order to build a model to cater our specific needs require a lot of trials and errors, especially our model is relatively new and still many on-going researches in the field.

2.5 Value of the Project

The proposed project would not directly bring monetary value to the company, because it is not a new product that can be sold separately, rather it is just an additional feature to the existing product. However, this feature might be able to help the company to gain a new market share, if it turns out that there is a market demand for this feature. Moreover, if the competitors have implemented a similar feature, this project could retain our users from switching, and hence there is a potential opportunity cost to the company. Lastly, the value of this project can grow over time, because GAN is still a relatively new idea in the research, there are still room for creative application, and also deep learning in general is still in an early development stage in the business world.

3. Characterizing the Data and Data Processing

3.1 Four ‘V’s

Volume

Media contents like songs, images and videos are one of the most fastest growing and abundant data out there. Media data are constantly being uploaded to social medias. Not only the breath has increased, but also the width, for instance, these days our smartphone camera is getting bigger sensors, hence, the pictures taken are higher in resolution. In practice, neural networks architectures are usually wide and deep, so it requires larger amount of data to train, especially tasks like neuro-linguistic programming (NLP) and image classification. Nevertheless, in our case, in order to train the GAN model to be able to generate a specific style, we need data of the same genre. For example, if we were to train a GAN model to generate instrumental music, we will need the song tracks that are only contains instrumental music instead of songs with vocal sound. In Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Douglas Eck, Karen Simonyan, and Mohammad Norouzi (2019) paper, the researchers used 70,379 samples to train a GAN model that can produce high quality instrumental music. Thus, the high volume of media data is advantageous to our project.

Velocity

Although new songs and images are generated very frequently, we are not particularly concern with it. Unlike some projects such as news feed analysis, the model or statistical inferences needed to be updated constantly, but our project, the model does not update frequently. Therefore, regardless of the velocity of the data, it has little to none impact to our project.

Variety

The variety of big data can be beneficial for problems like activity recognition and context recognition. However, there are progresses and researches being made in multi-modal learning, especially in deep learning, which allow us to better model different variety of data. In our case, there are yet to have any application or use case of different source of data, GAN models are usually trained on a single type of data.

Veracity

For audio data, it is usually very well produced. For example, the NSynth Dataset is stored in Musical Instrument Digital Interface (MIDI) file format, which directly records the instrument trigger instead of the sound produced by the instrument (J. Engel et al., 2019). Whereas for images, different devices capture different quality of images. Thus, in order to train our GAN model, we need to gather high quality images of the same style. However, GAN is an unsupervised generative model, hence the accuracy of the data is not the primary concern. While health diagnosis or monitoring data science projects, the veracity is important, and precise measurements are needed from the monitoring devices or sample data that represent the entire population.

3.2 Data processing

Data Storage

The project per se do not need any complex database system such as RDBMS, RDD, because we do not perform any task or query like sort, select, merge, join or group. Having said that, the company might have their own robust file system or databases system to store their music or photos. For example, Apple will have music management system for their Apple Music, and Google will have their photo management database system for their Google Photos.

Data Extraction

Prior to training our model, we need to extract the data from the database or file system, and subsequently extract features like RGB colour space and frequencies from our images and audio tracks respectively. Then it will be read into a tabular format like data frame and matrix for ease of training. There are many advance technics to extract features from image and audio data, they can be extracted into different representations for training, which will be discussed more in-dept in data analysis section.

Data Wrangling

As mentioned before, our data will be pretty well produced, but we still can do some basic wrangling like checking missing data, removing outliers, removing data with near zero variance, scaling, centering and removing variables with perfect linear combination.

4. Resources

4.1 Data

Open Data

In order to train our GAN models, we need two kinds of data: images, and audio tracks. There are numerous publicly available datasets we could use such as NSynth Dataset for audio tracks, and images data from ImageNet. ImageNet is an image database organized according to the WordNet hierarchy, hence it is useful for getting a specific type of images or labelled images for supervised training. While Magenta also offers some MIDI audio dataset like MAESTRO, Groove MIDI Dataset (GMD) and NSynth Dataset.

In-house data

As mentioned earlier, this project could dwell in companies like Apple or Google, and you could imagine these behemoth tech companies have huge amount of the data we need. For example, Google's Google Photos and Apple's iCloud have stored immense amount of photos. For audio tracks, both companies have their own music streaming service like Apple Music and Google Play Music. However, using data that are provided from users or artists might violate privacy or legal agreement, therefore, companies need to take precaution before using any of the in-house data.

Third-party data

Lastly, we can always purchase these data from data vendor. However, as mentioned earlier, this project per se will not help the company gain any monetary value, so the Data Businessperson must consider the budget, opportunity cost, and et cetera to make a valid justification to purchase these data.

4.2 Software

Wrangling and Visualization Tools

Python will be our primary software for building the models and wrangling the data. Python is the most popular programming language in the data science community, and there are good reasons for that. One of them is it has a huge number of statistical learning libraries, and some of them are maintained by big companies like Google (Tensorflow) and Facebook (Pytorch). For wrangling tasks, it has libraries like 'Numpy', 'Pandas', 'BeautifulSoup' and so on to read and manipulate various type of data such as 'csv', 'json' and 'html'. It also has some pretty good libraries on visualization like 'matplotlib', 'bokeh', 'seaborn', 'plotly' and et cetera.

Modelling and Production Tools

TensorFlow is perhaps the best end-to-end platform for building, training and deploying deep learning models from scratch. Once we train our GAN model in Python using TensorFlow package, then it can be imported to TensorFlow.js in JavaScript for web production, or TensorFlow lite for mobile production.

TensorFlow is a lower-level tool for building any deep learning model. Hence, for our project, we could use Magenta which is another platform that is specifically catered for making Music and Art using machine learning. Magenta is powered by TensorFlow and build by Google AI, it has wide variety of resources such studio, demos, blog posts, researches, pretrained-model and more that can help us to build our project. For instance, its Magenta Studio application has built-in music plug-ins for advanced music generation as shown below, these plug-ins are pretrained.

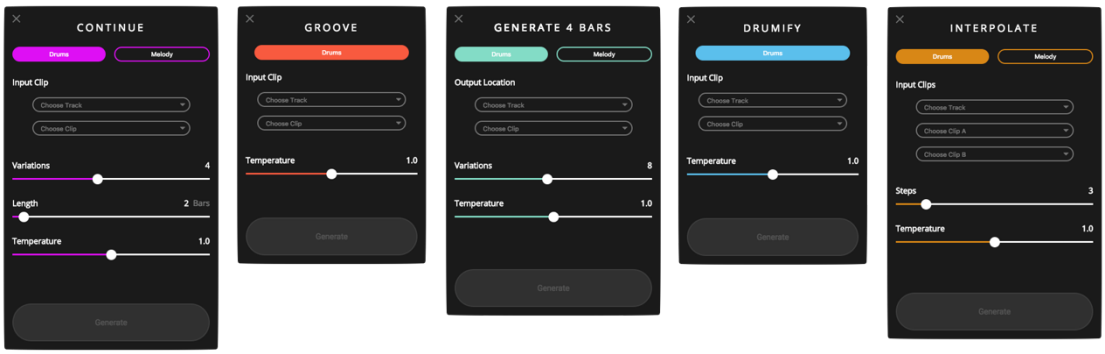

Image credit: https://magenta.tensorflow.org

Pretrained Model

Since our project does not involve in any predictive task or statistical inference, and all we need is the generative part of the GAN model. Hence, we can actually use the pretrained model that trained by others, and it can be found in Github or API provided in some blog posts. Generally neural networks with deep architecture are computationally expensive, therefore, using pretrained model can save us some time and it does not required any data. A few examples of pretrained models from Magenta to run on web can be found here: https://tensorflow.github.io/magenta-js/.

5. Data Analysis

5.1 Introduction

Although only GAN model has been discussed throughout the proposal, it is by no means the only model that can produce synthetic image or audio. For example, PixelCNN++ is one of the generative models that can generate high resolution images, and MuseNet for generating high quality audio. However, GAN model has some distinct training properties that enable it to generate sharper images, which will be discussed in this section.

5.2 Taxonomy

GAN is a class of artificial neural network (ANN). But unlike other ANN, GAN has two networks: generative network and discriminative network. In this case, we will only be using the generative part after we trained the model, so we will regard it as a generative model. However, the discriminative network can be used for other purposes like improving train time for other classification models, cyber security, adversarial attacks, and et cetera. Moreover, GAN is an unsupervised learning model, and it does not require labelled data, although the discriminative network is trained via supervised learning. As we can see GAN is a relatively special model, and it still in the stage of ongoing research.

5.3 Adversarial Training

What makes GAN special is its training process. As mentioned above, GAN has two networks to be optimized, and both of them are trying to achieve a different goal. Imagine there are two agents who are competing each other. One agent (generative network) is like the counterfeiter always trying to find a better way to deceive the another agent (discriminative network) who is the police, and the police will keep on improve their skills to better identify the counterfeit. After iterations of simultaneous training, they both reached a nash equilibrium, and the discriminative network is unable to identify whether the generated samples are real or fake. At this point, we have successfully trained our generative model. Nevertheless, this is an oversimplified analogy of how GAN are trained, in reality there are many problems such as non-convergence, vanishing gradients, mode collapse, minimax game, non-saturating loss function and many more to consider (Goodfellow, 2014). Therefore, the data researcher in this project will need to understand concepts like mini-batch normalization, double feedback loop learning, label smoothing and heuristic non-saturating cost, so that they are able to formulate the loss function and create the training process.

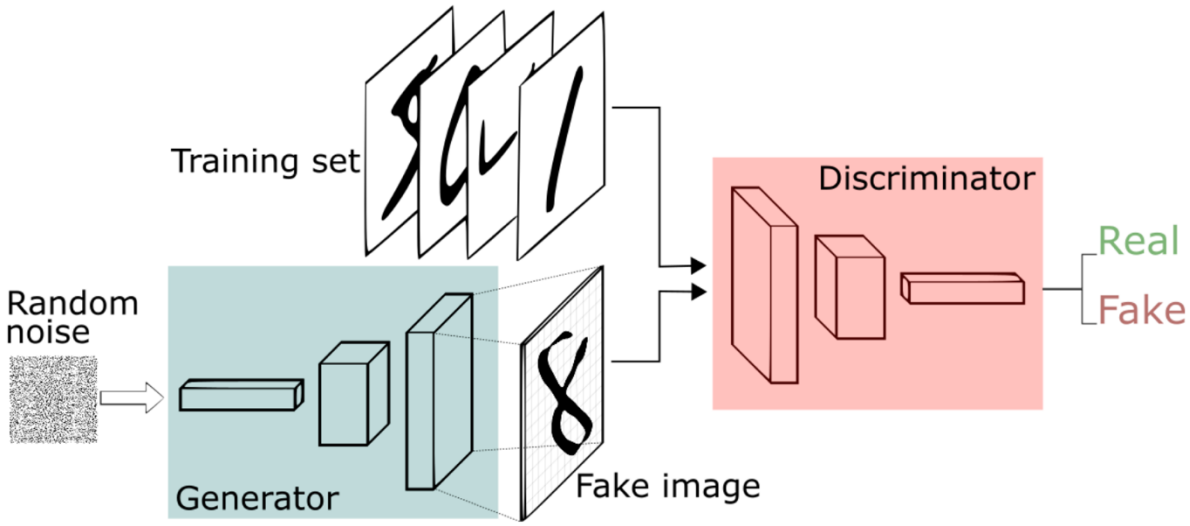

Image credit : Thalles Silva

5. 4 Implementation

Network Architecture

For our project, we need to design the neural networks architecture differently for both audio and image because of their differences in input representations, output, neural network layers, and et cetera. For audio, the input can be in spectral representation such as short-time Fourier transform (STFT) and angle phase (J. Engel et al., 2019). Whereas for image, the inputs can be represented with RGB colour space for each pixel. The generative network for both can be convolutional network which could involves striding, upscaling, normalizing, padding, inversing and transposing. The exact design of these neural network layers is again decided by the data researchers.

Latent Vector Space

In order for the user to control a certain aspect of the generating process such as style and mood of the music or image, we need to understand the latent vector space of the neural network. The latent vector space has arithmetic properties that allow us to manipulate the output, for example, we can define ‘man’ - ‘man’ + ‘sunglasses’, and we get an output of a woman wearing sunglasses. Hence, the data researcher and data creative can work together to design these controls for the user.

5.5 Conclusion

As we can see different data science project requires different focus during data analysis. For instance, credit risk or medical-related data science projects focus on things like statistic inference, classification, predictive models, accuracy, precision and recall. But our project does not concern any of that, instead we are focusing on building a generative model that to create synthetic samples that are indistinguishable to the human eyes or ears.

Reference

Engel, J., Agrawal, K. K., Chen, S., Gulrajani, I., Donahue, C., & Roberts, A. (2019). Gansynth: Adversarial neural audio synthesis. arXiv preprint arXiv:1902.08710.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. In Advances in neural information processing systems (pp. 2672-2680).

Harvard Business Review. (2012). What a Big-Data Business Model Looks Like. Retrieved 2018-08-25, from https://hbr.org/2012/12/what-a-big-data-business-model.

NIST. (2015). NIST Big Data Interoperability Framework: Volume 6, Reference Architecture ,1500(6).