Using GANs to Generate Synthetic Data

Joseph Despres and Yunus Shariff
Michigan State University

MICHIGAN STATE UNIVERSITY

Introduction and Motivation

The aim of this project is to extend the Generative Adversarial Network beyond image generation into generating synthetic data. The objective of any GAN is to train a Neural Network to recognize a fake, while training an adjacent Neural Network to spot the fake, the model has converged to a desirable solution where fake data are indistinguishable from the real [1]. This area has shown substantial progress in the past several years today's GANs are known to generate astonishingly accurate generating images. Applications for this technology have yet to be fully realized. This project explores one of many potential use cases.

Potential Use Cases

Although data are more and more available in seeming abundance, there is still a need to generate more. First, big datasets are often convenient to collect and not controlled.

- **Use to generate additional datapoints in Controlled experiments**

Since These data will be generated, confidentially is less of a concern. Therefore, these data could be used by peole training models outside of a pariticular organization.

- **Use to generate data to fit models where data are confidential**

Method

The challenges associated with training a GAN to generate synthetic data is a row in a data frame is several orders of magnitude smaller than a full image. An artificial nerual network has no difficulty fitting to it. The challenge is getting the right fit. This is a ballance between fitting the data so well it mimics the data seen and underfitting where it cannot generate. Therefor where the process could be generative yet,

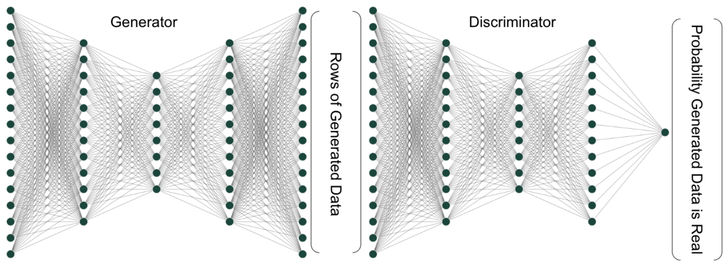

Figure 1: Basic GAN Architecture

Experiment

Here we will use a curitated dataset to generate synthetic copies.This project we use New York City Taxi from 2016[2]. After assembling the GAN archetecture, we will train a GAN on a subset of the data, then retrain adding more and more datapoints.

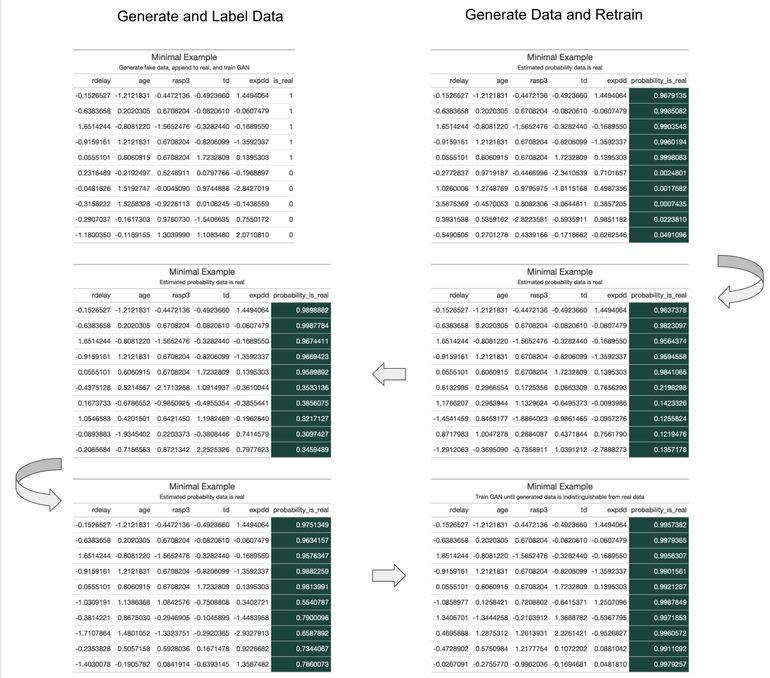

Figure 2: GAN Generating Synthetic Data

Results

Training this GAN optimizes a neural network to train a generator that is able to output data that in indistinguishable by discriminator. Notice In Figure 3, the GAN is able to fool the discriminator in training. This is on a log-scale because during this process the discriminator.

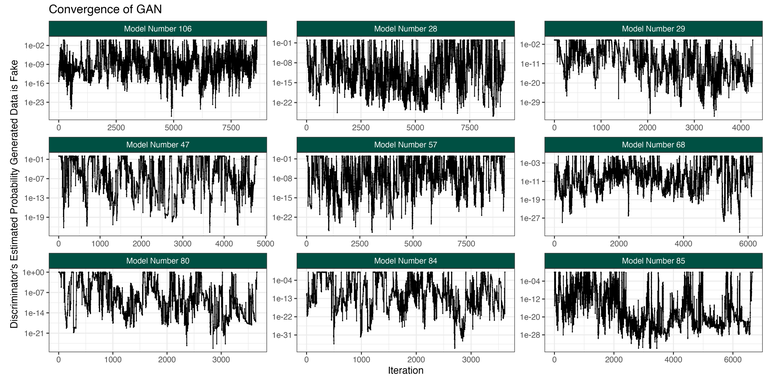

Figure 3: GAN Training Metrics

References

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[2] New York City Taxi and Limousine Commission.

.