INL/EXP-22-68329-Revision-0

Synthetic Data Generation Using Machine Learning

July 2022

Eduardo Antonio Trevino

Changing the World's Energy Future

INL is a U.S. Department of Energy National Laboratory operated by Battelle Energy Alliance, LLC

INL/EXP-22-68329-Revision-0

Synthetic Data Generation Using Machine Learning

Eduardo Antonio Trevino

July 2022

Idaho National Laboratory
Idaho Falls, Idaho 83415

http://www.inl.gov

Prepared for the
U.S. Department of Energy
Under DOE Idaho Operations Office
Contract DE-AC07-05ID14517

Synthetic Data Generation Using Machine Learning

Eduardo Trevino (Intern)

Shiloh Elliot and Ashley Shields (Mentors)

The University of Texas at San Antonio™

Idaho National Laboratory

Background

- Machine learning is a powerful, but complex technology. With complexity, comes a great benefit for **the flexibility of a personalized result**. Like any system, the confidence of a **machine learning model's ability to perform**, **increases**, as it becomes **personalized to fit** a particular problem. **New architectures**, and machine learning **techniques**, also **yield increasingly reliable models** so long as they are **molded to fit** the desired result.
- Therefore, when working on a solution using machine learning it is **crucial** for the team **to optimize** their models. Machine learning **optimization depends** on several components, but most importantly generating high confidence models requires, **personalized architectures**, and **reliable data**.

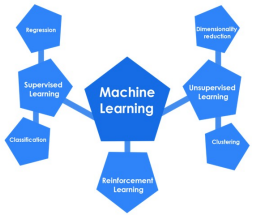

Fig.1: Complexity of machine learning and its various structures for personalized solutions IB (9 April 2018).

Problem

- When optimizing a machine learning model to fit your solution, a common **bottleneck** is the **scarcity of data**. For example, in my research, my team focused on detecting nuclear facilities using satellite imagery, however we had scarce data because of the limited number of nuclear facilities worldwide.
- Moreover, another common **data constraint** is **private data** which includes medical records, identification records, credit reports, mail and other forms of **personal information**. Machine learning models can prove to be a powerful disease detection tool for medical professionals, however training models on patient data **violates patient's medical privacy** and other legal complications.
- Furthermore, simulation scientists that build prediction models within machine learning require substantial data, generally the most within the scope of the field, as these models are used in finance, or safety, to make decisions that affect a large portion of our planet. Quickly, these scientists **run out of probable outcomes** and **maximize their data**. To expand on a common maximized dataset, an example is when dealing with the population. As it stands, according to the census bureau, there are currently 332,303,650 people in the United States, if a simulation scientist was taking one datapoint for every person, for example gender, the machine learning model's **data limit** for gender prediction would maximize after 332,303,650.
- Finally, synthetic data has a critical role in **providing equal opportunity** within the field of machine learning. As aforementioned, data is fuel to machine learning models, and because **certain parties have access to far more real user data** their models tend to be **unfairly powerful and advantageous** when compared to their startup competitors.

Method

- Most of machine learning consists of **supervised and self-supervised architectures**. Selection between them is **dependent on the problem**, with both models having their unique strengths. In my research, I found that a self-supervised model would be the best method in generating synthetic data. So, what is a self-supervised model? Difference? Similarities? A **self-supervised** model is a machine learning technique where the **model learns without** any **interference**. Naturally, a **supervised model** typically consists of an **analyst** in the background **guiding the model along** its training/learning session.

However, under the hood in the back-end **both models depend on a neural network**. A simple neural network's architecture consists of an **input layer, hidden layer, and an output layer**; with which the model algorithmically makes **decisions, predictions**, and **other** fundamental **analysis**. These **networks consists of nodes** which are **controlled** by positive or **excitatory**, and negative or **inhibitory weights**.

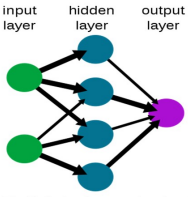

Fig.2: A simple neural network architecture. Hardesty, Larry (14 April 2017). "Explained: Neural networks". MIT News Office.

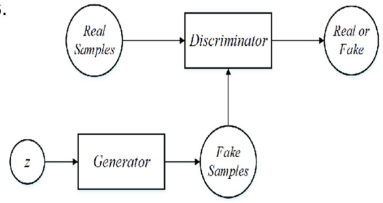

Fig.3: Flow of a generative adversarial network (GAN).
Goodfellow, I. J., "Generative Adversarial Networks", https://arxiv.org/abs/1406.2661, 2014

- Inside the realm of self-supervised architectures, I decided on the **Generative adversarial network** (GAN). A GAN **consists of 2 neural networks** competing in a **zero-sum game**. Given a training dataset the **discriminator** net **learns** on the **real dataset**, then **discerns** weather the data generated by the generator net is **real or fake**. Once the discriminator net is **fooled** by the **generator** net then we have an **efficient synthetic data generator**.

Results

At first the generator was generating colored noise. As the discriminator was trained for longer, the generator began to increasingly generate better data.

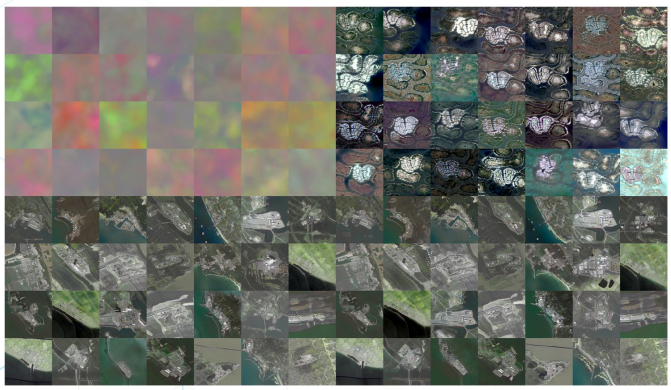

Fig.4: Evolution of a generative adversarial network (GAN) generating data.

Conclusions

- Machine learning requires a vast amount of **reliable data** and **testing**. With synthetic data we can **target the common bottlenecks** in systems, including data scarcity, private data, maximized data, and unfair data advantages.
- With plentiful datasets we can **stimulate** machine learning **systems** and **quantify our results with increased confidence**.

Acknowledgements

The author would like to acknowledge and thank the **National Nuclear Security Administration** (**NNSA**), and give a special thanks to **Shiloh Elliot, Matthew R. Kunz, Mark J. Schanfein, Gustavo A. Reyes, Mark J. Schanfein, and Ashley Shields** for guidance on this research.

www.inl.gov

Idaho National Laboratory