Real and Synthetic Data

As we have explored throughout this course, data is at the heart of machine learning. Oftentimes, data is the limiting factor and in turn dictates model performance. A scarcity in the amount of data points or number of features hinders the predictive power of our model. The solution seems obvious: collect more data, but this is not always feasible. So what can an ML engineer do?

Fortunately, with recent advances in deep generative models such as generative adversarial networks (GANs) and variational autoencoders (VAEs), very realistic synthetic data can be generated from a small number of known examples. This is achieved by approximately "learning" the data distribution and drawing random samples from the latent space of the generative model. This can be done on any type of data, allowing us to generate time series information (such as voice or sensor data), as well as photorealistic images. Using these new techniques, ML engineers can now bootstrap machine learning with machine learning -- that is, they can train models to generate more data with which to train other models!

What is Synthetic Data?

Imagine a power plant engineer that wants to use data to prevent a nuclear meltdown. They clearly cannot generate sample nuclear meltdowns to obtain relevant data, but they still need to be able to somehow make predictions. So what can they do? Well, if they have a very small amount of data from previous historical meltdowns or a simulation, then they can possibly use one of these techniques to train a model to generate enough data such that it becomes feasible to train an anomaly detection classifier using supervised learning. In this case, synthetic data is the only viable path forward.

The need to generate anomalous data is a commonplace situation in anomaly detection applications, since often the so-called “anomalies” are undesirable and can be extremely costly or damaging. Perhaps two of the most important applications of anomaly detection are for industrial processes and medical devices, where it can be used for predictive maintenance (fixing something before catastrophic failure or death).

Predictive maintenance will likely become an important feature of Industry 4.0, reducing costs by predicting component failures in advance and resolving the issue before it becomes more severe. An open-source dataset was created for this aiding with this specific application, referred to as MIMII (Malfunctioning Industrial Machine Investigation and Inspection). The MIMII dataset focuses on sounds exhibited by various components and can be used as a starting point to generate synthetic anomalous data for a variety of industrial components, reducing the need for individuals or organizations to collect these data.

Similarly, in the medical device space, anomaly detection allows for moderately malfunctioning devices to be detected on and replaced before they pose life-threatening issues to the device wearer. The ECG5000 dataset was created for this purpose, and provides 20 hours of ECG data - a type of data used to study heart health. Anomalous data is also provided in this dataset which corresponds to an individual with severe congestive heart failure. The dataset is open-source and can be freely used by anyone working on medical devices, and can be used as a starting point to train an anomaly detection model or to create more anomalous data.

This may sound incredible - we can effectively train models by simulating new data based on our current data. However, you have already done this in this course! When you use Data Augmentation, you are effectively expanding your dataset with synthetic data! That said, there are concerns when relying heavily on synthetic data that should be highlighted.

Quality Concerns of Synthetic Data

Real-life datasets are often complex and multifaceted. For example, when recording sounds, the actual sound plays an important role, but so does the microphone, background noise, microphone orientation, as well as myriad other factors. Creating synthetic data that is able to mimic this variability is infeasible, and currently there is no universally accepted method for generating high-quality synthetic data. One important reason for this infeasibility is that data must often go through multiple preprocessing stages before the generation of synthetic data. During this preprocessing stage, assumptions are implicitly made about the data which directly influence the synthetic data, meaning it is not purely based on the properties of the unprocessed data.

Similarly, if only a small amount of real-life data is collected, synthetic data generated using it may exhibit a high degree of bias. This can make it difficult to effectively extrapolate to other environments. Thus, it is still necessary to have access to as large and diverse a dataset as is possible given the contextual constraints to improve the generalizability of our synthetic data. The quality of the generated synthetic data must also be assessed, which can be problematic since the “quality” of a dataset can sometimes be very complex or specific, and not readily described by standard metrics.