Introduction

Hello you all. My name is Anton Savelyev, I am a student of group 107M-22 and today I would like to tell you about Synthetic Data Generation for Machine Learning

Abstract

Data is one of today's most valuable resources. But collecting real data is not always an option due to the cost, sensitivity, and processing time. But synthetic data can be a good alternative to rely on for training machine learning models.

What is synthetic data?

Synthetic data is artificial data that mimics real-world observations and is used to train machine learning models when actual data is difficult or expensive to get.

Synthetic data by itself does not have to be only computer generated. If we think about human history way before computerization, synthetic data existed as well, but it was only generated by humans. Draws for example.

Types of synthetic data related to its composition

Regarding its composition, there are two types of synthetic data: partial and full.
The **partial type** is a data set that includes synthetic data and real data from existing observations or measurements. An example can be a generated image of a car inserted in a photo of a real setting.

The **full type** refers to data sets with only synthetic data. An example would be a generated image of a car in a simulated environment. When choosing whether the data set is going to be fully or partly synthetic, the decision should depend on the main purpose. For instance, fully synthetic data gives more control over the data set.

When and why synthetic data is used?

Synthetic data can be used for many purposes. In fact, synthesized data can serve basically any goal of any project that would need a computer simulation to predict or analyze real events. There are several key reasons, a business may consider using synthetic data.

Besides to the obvious **Cost and time efficiency** there are cases in which data are rare or dangerous to accumulate. Dangerous real data could be exemplified by road accidents that self-driving vehicles must react to. In that case, we can substitute synthesized accidents.

Another advantage of synthetic data is **Easy labeling and control.** Using the example of **Image and video data generation** It's like we're doing the opposite: instead of labeling existing areas, we procedurally draw them as a mask, which then can be interpreted as a label.

Types of synthetic data

In addition to **Image and video data**, other types of synthetic data are used in business

Tabular data generation

Usually, tabular data includes way more sensitive and private details than other types. Exactly for these reasons, it must not just be anonymized but synthesized. Anonymization of data requires removal from data set characteristics by which a person can be identified. But this process is tricky.

Text data generation

Text and sound synthetic data have less frequent use in business, with greater use in research and art projects. Yet, textual data can be used to train for example chatbots, algorithms that check email boxes for spam, or machine learning models that detect abuse.

Cases of synthetic data application

Despite its novelty, this methodology is already in full use.

Self-driving vehicles

One of the cases is learning self-drive vehicles by Waymo that uses synthetic data to **train**. Waymo has created its own deep recurrent neural network and In this environment, a vehicle is trained on labeled synthetic data along with real data to drive safely, while recognizing objects and following traffic rules. Simulation imitates both good and bad road situations for self-driving cars to learn.

Virtual factory facility

Another case is **Virtual factory space** for robot navigation and production simulation for BMW designed by NVIDIA. The aim of the simulation was to prototype a future factory facility, creating its digital twin. In the blended reality, BMW's global teams were able to collaborate using various software packages to test, configure, and optimize the virtual production. Synthetic data was used to train models for robot movements within the virtual environment of the BMW factory.

Conclusion

Real data is extremely valuable but sometimes there are limitations in its processing such as privacy issues, and sometimes it becomes too precious, which is quickly reflected in the time and cost of obtaining it. And this is when synthetic data can save the day: being generated faster, cheaper, and in a fully controlled environment if needed.