SCIENCE · PASSION · TECHNOLOGY

Open Thesis / Project

Privacy Analysis of LLM-based Tabular Data / Sensor Data Generators

Embedded Learning and Sensing Systems Group

Motivation

The success of large-language models (LLMs), especially with regard to their generative capacities, made them interesting for applications outside of natural language processing (NLP) tasks. One such task is to generate synthetic tabular data, where LLMs outperform the state-of-the-art. However, it is unclear how this method performs with respect to privacy metrics. Recent literature has shown that Diffusion Models, a different kind of large generative model, most famously used in image generators such as Stable Diffusion or DALL-E 2, are much less private than previous generative models such as Generative Adversarial Networks (GANs).
Since tabular data is often used in domains where privacy is crucial (e.g. medicine, finance or IoT sensor data), it is critical to fill this gap and develop an understanding of how well LLMs perform with respect to privacy of the training data.

Interested? Please contact us for more details!

Target Group

Students in ICE, Computer Science or Software Engineering.

Thesis Type

Master Project / Master Thesis.

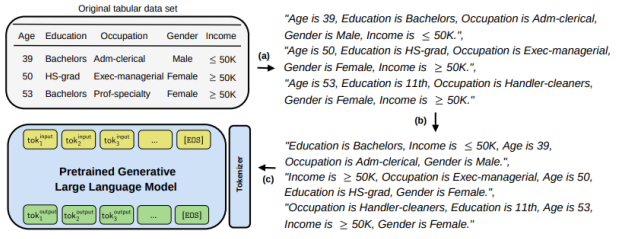

Image source: https://tinyurl.com/5ft42y8w

Goals and Tasks

The goal of this work is to generate synthetic data using LLM-based methods. First experiments should be done using standard tabular datasets (e.g. diabetes or adult dataset), later on more challenging IoT sensor data.

- Thorough literature research on the topic;
- Select suitable privacy metrics;
- Familiarize yourself with the synthcity https://github.com/vanderschaarlab/synthcity/ and GReaT https://github.com/kathrinse/be_great frameworks
- Design and run experiments to investigate privacy in tabular data from LLMs;
- Summarize the results in a written report

Requirements:

- Creativity, interest in state-of-the-art machine learning methods;
- Programming skills;
- Knowledge of in deep learning frameworks (TensorFlow/Keras or PyTorch) is recommended, but not mandatory.

Used Tools & Equipment

- A laptop
- Sufficient computational resources (will be provided as needed)
- Your skills

Contact Persons

- Franz Papst (papst@tugraz.at)
- Dr. Olga Saukh (saukh@tugraz.at)

Institute for Technical Informatics
Embedded Learning and Sensing Systems Group