Project Proposal

Team name: BlackBox

Project Title: Study of Adversarial Attacks on Privacy in Large Language Models

Project Summary:

State-of-the-art large language models (LLM) are trained on vast swathes of data from the internet and possibly private sources that contain personally-identifiable and sensitive information. In the future, we may anticipate there to be domain-specific language models that are trained, for example using medical transcription data, to provide healthcare specialists with a tool to generate differential diagnoses given a list of symptoms from the patient. Adversarial parties may attempt to exploit weaknesses in these models to extract sensitive information from the training data. This project aims to understand how effective privacy-preserving training methods such as differentially private stochastic gradient descent (DP-SGD) are at preventing these exploits.

Approach:

We plan to build a generative neural network (GNN), train it on a natural language dataset that contains sensitive information, and then conduct a series of experiments. The experiments will focus on how DP-SGD, and possibly other techniques, can be used to mitigate adversarial attacks aimed at getting the model to output sensitive information contained in the training data.

First, we will obtain and pre-process a dataset that contains sensitive information. The Medical Transcriptions is a publicly available dataset with clinical data that can be augmented with mock identifiers, including name, addresses, birth date etc. The pre-processing will involve tokenization, removing special characters and stopwords, and converting identifying fields to a standardized format.

Second, we will build our own architecture as a starting point, drawing inspiration from state-of-the-art models like GPT-2, and tailor it as needed. We will train the model on the pre-processed medical transcription dataset, monitoring the loss and validation metrics while tuning the hyperparameters to ensure convergence.

Third, we will demonstrate an exploitation of our model by crafting queries that extract personal information from the training data. We will evaluate the model's responses to these queries, highlighting instances where sensitive information is leaked.

Next, we will familiarise ourselves with the concepts of Differential Privacy and the DP-SGD algorithm. We will modify the training process to use DP-SGD instead of regular SGD, adjust the privacy budget and noise parameters, and then re-train the model using the modified training process.

Last, we will re-run the privacy leakage demonstration on the DP-SGD-trained model. We will compare the results to the original model to assess the reduction in sensitive information leakage. We will analyse the trade-off between privacy protection and model performance. As a stretch goal, we will conduct further experimentation to search and quantify any drawbacks in the DP-SGD training approach and potentially propose modifications or a novel mechanism as an improvement.

Resources / Related Work & Papers:

[1] “A Study on Extracting Named Entities from Fine-tuned vs. Differentially Private Fine-tuned BERT Models”, Diera et al.

[2] “Deep Learning with Differential Privacy”, Abadi et al.

[3] “Calibrating Noise to Sensitivity in Private Data Analysis”, Dwork et al.

[4] “Text Processing Like Humans Do: Visually Attacking and Shielding NLP Systems”, Eger et al.

[5] “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, Devlin et al.

[6] “Language Models are Unsupervised Multitask Learners”, Radford et al.

[7] “Differential Privacy Series Part 1 | DP-SGD Algorithm Explained”, Davide Testuggine and Ilya Mironov.
[https://medium.com/pytorch/differential-privacy-series-part-1-dp-sgd-algorithm-explained-12512c3959a3]

[8] “Privacy-Preserving Generative Deep Neural Networks Support Clinical Data Sharing”, Beaulieu-Jones et al.

Datasets:

Medical transcription data scraped from MTSamples.com

Team members

Poorna Nataranjan

Stephen Valenta

Jenny Yang