Application of Selective Differential Privacy (SDP) for LLM Privacy Protection

Pals Chinnakannan, Naveed ul Islam, Madhukar Yedulapuram

Introduction

The rapid growth in Generative Artificial Intelligence around Large Language Models (LLMs), and the advancement in the associated techniques for fine-tuning and prompt tuning have spurred many organizations to customize LLMs for solving business critical problems and for competitive advantage. This growth has resulted in deploying LLMs fine-tuned or prompt-tuned using a wide-range of data containing sensitive Privacy information. However, surveys and studies on the security of LLMs and associated privacy indicate potential scope for vulnerabilities and wide-spread privacy violation on the horizon. These concerns are clearly brought out by Yao et al in [1]. Differential privacy is a technique used to protect PII and other sensitive information through fuzzing such information in a dataset. The essence of differential privacy lies in the injection of "noise" to obscure privacy information present in datasets used in domains like LLM Applications, thereby preserving anonymity while maintaining the overall integrity and the utility of the dataset. Several different Differential Privacy Techniques have been proposed, researched and developed [2], [3]. “Selective Differential Privacy for Large Language Models”, and “Just Fine-Tune Twice: Selective Differential Privacy for Large Language Models” are recent additions to Differential Privacy. SDP is very appropriate for the sparse amount of privacy-related information present in very large datasets like those used in LLM Training and Fine-Tuning. This project proposes to evaluate the Selective Differential Privacy technique as proposed in the paper by W.Shi et al., [5], first by validating the claims in the paper and then by applying those techniques to Meta LLama-2-7b-hf, using the datasets identified in the paper.

Literature Survey

The fundamental Differential Privacy concept, the associated mathematics and the technology was pioneered by Dwork et al in their work on Algorithmic Foundations for Differential privacy [1]. The earlier work from M.Abadi and et al [2], applies the differential privacy concepts and technology to LLMs. These two research works serve as a good starting point for this project. In addition, a majority of the authors of this proposal researched Differential Privacy techniques in an earlier course work, which led to their seminal work on “Applied Differential Privacy on Large Language Models”. In that study and paper the authors have clearly brought out the privacy concerns of LLMs and have shown the revealing nature of LLM, specifically, LLama-2. This work is continuing in the same vein as Privacy of LLMs. Finally, the papers on SDM [4] and [5] serve as the foundation for this project.

Objectives

The objectives of this project is as follows:

- Application of the earlier effort of fine-tuning and prompt tuning of LLama-2-7b to this “SDP for LLM Privacy Protection project”, Specifically, reuse the infrastructure and environment to jump start this work.
- Build the foundation required for applying SDP technology on selected data sets.
- Procure and study the data sets described in the paper, specifically, 1) GLUE (Wang et al., 2018) a widely- used multi-task benchmark dataset for NLU, which contains sensitive information such as name and date. 2) Wikitext-2 (Merity et al., 2017) that contains Wikipedia articles with private information such as name and date. *3)* ABCD (Chen et al., 2021), a human-human customer service dialogue dataset under real-world scenarios with user private information such as name and order IDs.
- Apply the SDP principles on the LLMs described in the paper, such as 1) Roberta (Liu et al., 2019) for the NLU classification task and 2) GPT2 (Radford et al., 2019) for the language generation task and compare the results with the findings in the paper.
- Finally, apply SDP on LLama-2 using the above datasets and evaluate their performance.

References

[1] Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, Yue Zhang, 2024, A Survey on Large Language Model (LLM) Security and Privacy: The Good, the Bad, and the Ugly, https://doi.org/10.48550/arXiv.2312.02003

[2] Cynthia Dwork, Aaron Roth, et al. 2014. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3- 4):211–407.

[3] Martin Abadi, Andy Chu, Ian Goodfellow, H Bren- dan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security, pages 308–318.*

[4] Weiyan Shi, Aiqi Cui, Evan Li, Ruoxi Jia, Zhou Yu, 2022, Selective Differential Privacy for Language Modeling, https://doi.org/10.48550/arXiv.2108.12944

[5] Weiyan Shi, Ryan Shea, Si Chen, Chiyuan Zhang, Ruoxi Jia, Zhou Yu , 2022, Just Fine Tune Twice, Selective Differential Privacy for Large Language Models https://ar5iv.labs.arxiv.org/html/2204.07667