MIA and ADePT

1 Introduction

This report includes a brief introduction to key membership inference attack (MIA) concepts and a review of an influential study by Krishna et al. (2021).

2 Membership inference attack

Understanding how machine learning models leak information from training datasets is crucial for advancing privacy-preserving algorithms. Homer et al. first introduced the concept of membership inference attacks (MIAs) by proposing using published statistics from a genomics dataset to infer the presence of a particular genome within the dataset (Homer et al., 2008; Shokri et al., 2017; Tabassi et al., 2019). Shokri et al. were the first to introduce the MIA technique in the context of ML. They defined the fundamental question of membership inference as follows: *given a machine learning model and a record, determine whether this record was part of the model's training dataset or not* (Shokri et al., 2017). To answer this question, they devise a setup consisting of creating multiple "shadow models" and an attack model consisting of the following steps:

1. Create shadow model(s) that have a similar architecture and are trained on datasets resembling those used for the target model.

2. Test membership/non-membership examples on the shadow model to generate probability vectors for each class.

3. Train the attack model on the concatenation of the record and the shadow model prediction vector on *in* and *out* membership status.

The attack model learns to recognize the differences in the target model's behaviors on members and non-member data points and uses them to distinguish membership/non-membership based on the target model's output patterns (Shokri et al., 2017).

3 Overview of paper ADePT: Auto-encoder-based Differentially Private Text Transformation

3.1 Motivation

Differential privacy (DP) is a standard privacy protection technology that provides rigorous privacy guarantees for data used in statistical model training (Zhao and Chen, 2022; Hu et al., 2023). ADePT is an algorithm designed as a DP mechanism to transform texts for NLP tasks. Its theoretical privacy guarantee is verified through proof based on Dwork et al. (2009, 2014).

3.2 Component

ADePT (Adversarial Deep Privacy-preserving Text) consists of autoencoders and decoders. The autoencoders are text-based neural networks (such as LSTM). To transform the original texts, operations like clipping and noising are added to the latent sentence representations returned by the autoencoders. The decoders are used for text generation.

3.3 Evaluation

The success rate of MIA is used to quantify the effectiveness of the ADePT mechanism. The evaluation is motivated by two dimensions: first, if an attack model cannot accurately infer membership status, it indicates that the DP mechanism effectively preserves privacy by transforming the texts. This is assessed using the AUC-ROC, where a lower value signifies better privacy preservation. Second, if an intent classifier (IC) trained on DP-transformed texts achieves high performance, it demonstrates that the DP mechanism successfully retains utility, which is assessed by accuracy.

The evaluation results show that in general, ADePT mechanism outperforms the baseline

Redactive mechanism, obtaining lower AUC and higher IC accuracy. In addition, when utilizing Gaussian noise, ADePT achieves the best performance. Nevertheless, Krishna et al. also note that the encoders are too sensitive to additional clipping and noise, which decreases the utility of the mechanism.

3.4 Strength and weakness

Original	what are the flights on january first 1992 from boston to san francisco	show all flights boston to any time
ADePT	what are the flights on thursday going from dallas to san francisco	show all flights flights flights boston to any time

Table 1: Table from Krishna et al. (2021): example of a good and a bad output

In the experiment, ADePT is shown to provide DP and retains better utility than the baseline mechanism, mitigating the risk of adversaries inferring. By adding noise and applying clipping to latent representations, ADePT anonymizes sensitive details while retaining semantic utility. The manipulation of latent sentence representations is an innovative approach. The privacy and utility trade-off can be refined to meet specific requirements, offering flexibility.

However, as we can see from Table 1, although ADePT preserves syntactic structures, altered sentences lack natural flow. The privacy and utility trade-off leads to either a loss of semantic information, or degradation of privacy preservation. The mechanism may perform well in some classification tasks, but there is no guarantee of its effectiveness when generalized to other tasks, languages, or domains – particularly in NLG tasks and high-stakes domains. Finally, ADePT relies solely on AUC-ROC to measure its capabilities in privacy preservation, providing limited insight and interpretability.

References

Cynthia Dwork, Moni Naor, Omer Reingold, Guy N Rothblum, and Salil Vadhan. 2009. On the complexity of differentially private data release: efficient algorithms and hardness results. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pages 381–390.

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.

Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V Pearson, Dietrich A Stephan, Stanley F Nelson, and David W Craig. 2008. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS genetics*, 4(8):e1000167.

Lijie Hu, Ivan Habernal, Lei Shen, and Di Wang. 2023. Differentially private natural language models: Recent advances and future directions. *arXiv preprint arXiv:2301.09112*.

Satyapriya Krishna, Rahul Gupta, and Christophe Dupuy. 2021. Adept: Auto-encoder based differentially private text transformation. *arXiv preprint arXiv:2102.01502*.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE.

Elham Tabassi, Kevin J Burns, Michael Hadjimichael, Andres D Molina-Markham, and Julian T Sexton. 2019. A taxonomy and terminology of adversarial machine learning. *NIST IR*, 2019:1–29.

Ying Zhao and Jinjun Chen. 2022. A survey on differential privacy for unstructured data content. *ACM Computing Surveys (CSUR)*, 54(10s):1–28.