SeqPATE: Differentially Private Text Generation via Knowledge Distillation

Zhiliang Tian[1]*, **Yingxiu Zhao**[2], **Ziyue Huang**[2], **Yu-Xiang Wang** [3], **Nevin L. Zhang** [2], **He He** [4]
[1] National University of Defense Technology,
[2] The Hong Kong University of Science and Technology,
[3] UC Santa Barbara,
[4] New York University
tianzhilianghit@gmail.com, yzhaocx@connect.ust.hk, zyhuang94@gmail.com,
yuxiangw@cs.ucsb.edu, lzhang@cse.ust.hk, hhe@nyu.edu

Abstract

Protecting the privacy of user data is crucial for text generation models, which can leak sensitive information during generation. Differentially private (DP) learning methods provide guarantees against identifying the existence of a training sample from model outputs. PATE is a recent DP learning algorithm that achieves high utility with strong privacy protection on training samples. However, text generation models output tokens sequentially in a large output space; the classic PATE algorithm is not customized for this setting. Furthermore, PATE works well to protect sample-level privacy, but is not designed to protect phrases in samples. In this paper, we propose SeqPATE, an extension of PATE to text generation that protects the privacy of individual training samples and sensitive phrases in training data. To adapt PATE to text generation, we generate pseudo-contexts and reduce the sequence generation problem to a next-word prediction problem. To handle the large output space, we propose a candidate filtering strategy to dynamically reduce the output space, and refine the teacher aggregation of PATE to avoid low agreement due to voting for a large number of candidates. To further reduce privacy losses, we use knowledge distillation to reduce the number of teacher queries. The experiments verify the effectiveness of SeqPATE in protecting both training samples and sensitive phrases.

1 Introduction

Recent work has shown that sensitive user information in training corpora, such as addresses and names, can be extracted from text generation models [6]. Providing privacy guarantees to the training corpora of text generation models has become a critical problem. Differential privacy (DP) provides provable guarantees against detecting individuals in datasets. Deep learning models with DP guarantees ensure that the existence of a specific training sample cannot be detected.

NoisySGD [42, 3, 1] is a popular DP algorithm for deep learning that adds noise to the gradients. PATE [31] is another type of DP learning algorithm that transfers knowledge from teachers trained on private data to a student model, where noises are added to teacher predictions to satisfy DP. PATE is model-agnostic, and its privacy cost derives from the knowledge distillation process instead of the model gradients in NoisySGD [42, 24]. Therefore, the noises required by PATE do not scale with model size. Given this benefit, PATE has great potential for text generation, since large language

*This paper was partially done when Zhiliang Tian was a Ph.D. student at HKUST and a visiting scholar at NYU.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).