A COMPARATIVE ANALYSIS OF DIFFERENTIALLY PRIVATE PROMPT TUNING METHODS: PROMPTDPSGD, PROMPTPATE, AND DP-OPT

Pouria Dadkhah
pouria.dadkhah@gmail.com

1 INTRODUCTION

Large Language Models (LLMs) have become essential tools in various applications, but their dependency on sensitive private information for tuning prompts raises significant privacy concerns. Two notable methods to address these concerns are PromptDPSGD and PromptPATE, introduced in the "Flocks of Stochastic Parrots" paper, and Differentially-Private Offsite Prompt Tuning (DP-OPT), presented in a separate study. This paper provides a detailed comparison of these methods, focusing on their main purposes, algorithms, applications, pros and cons, mathematical details, type of method, and evaluation results.

2 METHODS

2.1 METHOD SUMMARIES

PromptDPSGD aims to efficiently and privately learn soft prompts using Differentially Private Stochastic Gradient Descent (DPSGD) without modifying the underlying LLM's parameters. This method is designed to optimize prompt embeddings while ensuring differential privacy during the training process.

PromptPATE leverages the Private Aggregation of Teacher Ensembles (PATE) framework to create differentially private discrete prompts. The method uses multiple teacher models to predict labels for a public dataset, with a noisy aggregation process to ensure privacy.

DP-OPT seeks to enable privacy-preserving prompt tuning for LLMs hosted on cloud platforms. It focuses on generating discrete prompts on local client devices and transferring them to cloud models, ensuring data confidentiality and privacy through differential privacy mechanisms.

2.2 METHOD ALGORITHMS

For **PromptDPSGD**, the algorithm involves initializing soft prompt embeddings P_0, computing per-sample gradients $\nabla P_t \ell(L_P, x_i)$ for each $x_i \in B_t$, clipping gradients to have a maximum ℓ_2-norm c, adding Gaussian noise $N(0, \sigma^2 c^2 I)$ to the clipped gradients, updating the embeddings using the noisy gradients, and outputting the final embeddings P_T.

The **PromptPATE** algorithm includes training teacher models using disjoint subsets of private data, having teachers predict labels for a public dataset, performing a noisy majority vote using the Confident GNMAX algorithm to ensure differential privacy, using the noisy labels to train a student model or create discrete prompts, and selecting the best student prompt based on validation accuracy using labeled public data.

In **DP-OPT**, the steps are initializing with an initial prompt π_0 and an empty set Π, generating a modified dataset D' by passing the data through the local model with the initial prompt, generating private prompt π_n from D' using DP-EnsGen, adding π_n to Π, selecting the prompt $\hat{\pi}$ from Π that maximizes accuracy on a validation set D_{val} using a differentially private mechanism, and outputting the selected prompt $\hat{\pi}$.

2.3 MATHEMATICS DETAILS

For **PromptDPSGD**, Gradient Clipping: $g_t(x_i) \leftarrow \frac{\nabla P_t \ell(L_P, x_i)}{\max\left(1, \frac{\|\nabla P_t \ell(L_P, x_i)\|_2}{c}\right)}$; Noise Addition: $g_t \leftarrow \frac{1}{|B_t|} \sum_i g_t(x_i) + \mathcal{N}(0, \sigma^2 c^2 I)$; Update Rule: $P_{t+1} \leftarrow P_t - \eta_t g_t$.

In **PromptPATE**, Noisy Voting: $n_j(x) = \sum_{i=1}^{N} \text{votes}_{i,j}(x) + \mathcal{N}(0, \sigma^2)$; Final Label: label $=$ $\text{argmax}_j\,(n_j(x))$.

For **DP-OPT**, Differential Privacy involves the Exponential Mechanism: $\Pr[\text{DP-Argmax}_\epsilon(h) = j] \propto \exp(\epsilon h_j)$; Privacy cost growth: $\epsilon \sim \sqrt{m}\epsilon_0$.

3 COMPARISON

3.1 APPLICATIONS

PromptDPSGD is suitable for scenarios where access to model gradients is possible, making it ideal for soft prompt tuning in tasks like text classification. **PromptPATE** is designed for black-box settings where only discrete prompts can be used, making it ideal for scenarios involving commercial APIs with limited access to model internals. **DP-OPT** is applicable in environments requiring privacy-preserving prompt tuning, such as healthcare and legal domains, where sensitive data must remain confidential.

3.2 PROS AND CONS

PromptDPSGD optimizes fewer parameters than full model fine-tuning and does not require access to LLM parameters, providing strong privacy guarantees with minimal performance loss. However, it requires gradient access, which may not be available in all API setups, and has limited applicability due to dependency on soft prompts.

PromptPATE is suitable for black-box models and provides high utility with strong privacy guarantees through noisy aggregation, but it is computationally intensive due to multiple teacher models and requires significant public data for effective knowledge transfer.

DP-OPT keeps sensitive data local, preventing exposure to untrusted cloud services, and shows effective performance across various tasks, maintaining accuracy close to non-private methods. However, it involves computational complexity due to added steps for differential privacy and utility trade-offs under stricter privacy budgets.

3.3 EVALUATION RESULTS

PromptDPSGD achieved near non-private baseline accuracy on various NLP tasks (e.g., 92.31% on SST2 with $\epsilon = 8$). **PromptPATE** showed high performance in downstream tasks (e.g., 92.7% on SST2 with $\epsilon = 0.147, \delta = 10^{-6}$) and was effective even when public data differs from private data. **DP-OPT** demonstrated competitive accuracy on SST-2, Trec, Mpqa, and Disaster datasets, with strong performance in transferring prompts from smaller models like Vicuna-7b to larger models like DaVinci-003, and balanced utility and privacy with larger models mitigating some accuracy loss under strict privacy budgets.

4 CONCLUSION

PromptDPSGD, PromptPATE, and DP-OPT each provide unique strengths and trade-offs for privacy-preserving prompt tuning. PromptDPSGD and PromptPATE focus on leveraging gradient descent and teacher-student frameworks, respectively, while DP-OPT emphasizes local prompt generation and transferability to cloud models. Each method demonstrates effective performance with varying computational complexities and privacy guarantees, making them suitable for different privacy-sensitive applications.