Optimizing Inference Time in Private Learning of Large Language Models

Sonakshi Garg and Vicenç Torra

Umeå University, Umeå, Sweden {sgarg,vtorra}@cs.umu.se

Abstract. Large Language Models (LLMs) have demonstrated state-of-the-art performance across various applications. However, these models which consists of millions of parameters still face challenges due to their computational inefficiency during inference time. In many real-time tasks, the inference time of the tasks should be in the order of milliseconds to be useful. Deploying LLMs in real-time requires reducing inference time. One effective approach is model compression, which reduces the model's parameters. Also, ensuring data privacy is crucial when using LLMs due to their vulnerability to privacy attacks. To address this, LLMs should be trained with Differential Privacy (DP) on private data. Developing a compact, efficient, and private domain-specific language model is an active area of research. To facilitate the deployment of efficiently compressed models with DP while minimizing any degradation in model utility, we propose our framework, Task-Specific Knowledge Distillation with Differential Privacy. In our approach, we prioritize task-specific distillation that generally enhances the performance of downstream tasks unlike traditional methods. Additionally, we leverage transfer learning by utilizing pre-trained models trained on similar tasks, and then use these models for private fine-tuning. We also emphasized on the initialization of student models with pre-trained models from open domains. We demonstrate the effectiveness of our framework on the GLUE benchmark datasets, employing the BERT-base model as our teacher model, and utilizing BERT-tiny and DistilBERT models as student models. Our framework showcases comparable accuracy to non-private learning methods while also improving the accuracy of student models compared to existing baselines. Remarkably, we achieve this while reducing the parameters of student models by 95%.