Abhijit Mishra

SentinelLMs: Encrypted Input Adaptation and Fine-Tuning of Language Models for Private and Secure Inference

Abhijit Mishra, Mingda Li, Soham Deo

School of Information, University of Texas at Austin
{abhijitmishra, mingdali, soham.deo}@utexas.edu

Abstract

This paper addresses the privacy and security concerns associated with deep neural language models, which serve as crucial components in various modern AI-based applications. These models are often used after being pre-trained and fine-tuned for specific tasks, with deployment on servers accessed through the internet. However, this introduces two fundamental risks: (a) the transmission of user inputs to the server via the network gives rise to interception vulnerabilities, and (b) privacy concerns emerge as organizations that deploy such models store user data with restricted context. To address this, we propose a novel method to adapt and fine-tune transformer-based language models on passkey-encrypted user-specific text. The original pre-trained language model first undergoes a quick adaptation (without any further pre-training) with a series of irreversible transformations applied to the tokenizer and token embeddings. This enables the model to perform inference on encrypted inputs while preventing reverse engineering of text from model parameters and intermediate outputs. After adaptation, models are fine-tuned on encrypted versions of existing training datasets. Experimental evaluation employing adapted versions of renowned models (e.g., BERT, RoBERTa) across established benchmark English and multilingual datasets for text classification and sequence labeling shows that encrypted models achieve performance parity with their original counterparts. This serves to safeguard performance, privacy, and security cohesively.

Introduction

In today's technology landscape, language models play a crucial role in numerous text-based applications. The rise of powerful pre-trained models like BERT (Devlin et al. 2019), RoBERTa (Liu et al. 2019), and GPT (Radford et al. 2018) and their successors, followed by a massive outburst of generative large language models such as GPT-3 (Brown et al. 2020), PaLM (Chowdhery et al. 2022), LLaMa (Touvron et al. 2023) has led to a surge in innovative approaches and designs for large-scale language model pre-training. Characterized by an initial phase of broad pre-training, followed by targeted fine-tuning for specific contextual tasks, these models have found widespread utility across a diverse spectrum of applications (Otter, Medina, and Kalita 2020; Li et al. 2022) involving text classification, sequence labeling and text generation.

The widespread use of applications on server systems accessed through the internet has raised concerns about user privacy. Pre-trained language models are powerful tools, but their size and resource requirements make them challenging to use on personal devices. This leads to a situation where models are deployed on cloud-based servers and users send their text inputs over networks that can potentially be intercepted. For instance, sending text from users to servers using methods like HTTP GET/POST can expose the data to attacks known as *Man-in-the-Middle (MitM)* attacks (Callegati, Cerroni, and Ramilli 2009), where an attacker sneaks between the user and the server, gaining access to and potentially altering the data being sent. This can result in data theft and misuse by malicious individuals. Another common worry related to using fine-tuned models on servers is the risk of unauthorized logging and misuse of users' personal data the organizations deploying these models (O'Neill 2003; Jones, McCullough, and Richman 2005).

To deal with these problems, strong protective measures like encrypting data and models can be undertaken, involving potent one-way string encryption techniques such as the Secure Hash Algorithm (SHA) and passkey-based secure hash algorithms like *Blake* (Fernandes et al. 2015). These methods convert input data into fixed-length hash values, which are incredibly difficult to reverse-engineer into the original input. However, encrypting input text introduces challenges for language models, typically trained and fine-tuned on plain text. This is due to two key reasons. Firstly, models pre-trained and fine-tuned on plain text may not adequately recognize encrypted inputs and may interpret them as *out-of-vocabulary* (OOV) terms. Secondly, encryption hash algorithms provide enhanced security mostly by transforming text in a way that sacrifices a significant amount of linguistic information. For instance, consider the words *cat* and *cats* in plain text. They are closely related, sharing morphology and lexical semantics. However, their respective encrypted forms using a 256-bit SHA encryption, *77af778b* and *d936608b* (truncated for brevity), bear no clear relationship. This poses difficulties in suitably modeling language tasks with encrypted inputs. Pre-trained language models thrive on linguistic patterns in the input text, a task made challenging by the transformation introduced during encryp-

Mishra, A., Li, M., & Deo, S. (2024). SentinelLMs: Encrypted input adaptation and fine-tuning of language models for private and secure inference. Proceedings of the AAAI Conference on Artificial Intelligence. 38(19), 21403–21411.