Label Privacy in Vertical Federated Learning with Parameter-Efficient Finetuning

Kseniya Shastakova
Scientific adviser - Aleksandr Beznosikov

Moscow Institute of Physics and Technology

shestakova.ko@phystech.edu

March 19, 2024

Table of contents

High-level problem statement

We study minimization problem

$$\min_x f(x)$$

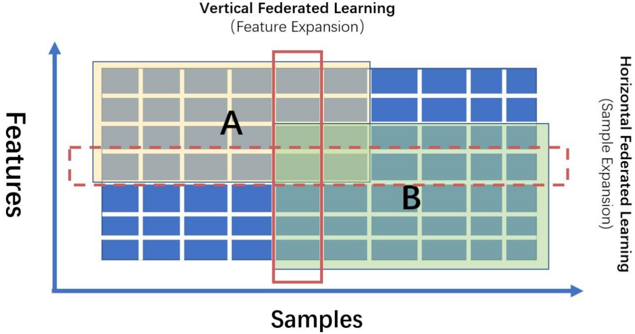
where f is usually a loss function and x is a model parameter

Introduction to VFL

Federated Learning (FL) - several parties (a server and clients) collaboratively train a ML model

Horizontal Federated Learning (HFL) - different clients have different data samples

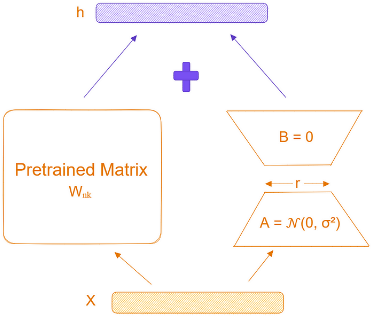
Vertical Federated Learning (VFL) - different clients have different features describing same samples



Introduction to PEFT

Parameter-Efficient Fine-Tuning (PEFT) - training a relatively small number of additional weights to adapt a pre-trained model to a specific task

Low-Rank Adaptation - adjusting weight matrix W by adding $\delta W = AB$, where A and B are matrices with low ranks



VFL components in our case

- Smaller model ("head") with it's local dataset, which aims to minimize loss-function avoiding data leakage

$$\min_{\theta} L(h(x, \theta))$$

- Large model with API:
  forward(x, θ) - computes model activation with input x and adapter parameters θ
  backprop(x, θ, g_h) - receives gradients of an arbitrary loss function L w.r.t. model activations $g_h = \frac{\partial L(h(x,\theta))}{\partial h}$ and returns the gradients w.r.t adapter parameters $g_\theta = \frac{\partial L(h(x,\theta))}{\partial \theta}$

Motivation

- Large models like ChatGPT are getting more and more popular
- It would be great to adapt them for your specific goals ~~imagine you have CS-oriented ChatGPT for solving your homework~~
- When fine-tuning a huge model, one still wants to keep one's data private

Related work

Out research aims to expand the results from [1], that shows that activations and gradients can be used to predict labels with k-means algorithm and describes possible defense based on:

- linearity of **backprop**(x, θ, g_h) w.r.t. g_h, thus it is possible to split the gradient $g_h = \sum_{i=1}^{m} \alpha_m \hat{g}_h^{(m)}$
- splitting the LoRa parameter θ into $\theta_1, \ldots, \theta_n$ and using $h'(x, \theta_1, \ldots, \theta_n) = \sum_{i=1}^{n} W_i \odot h(x, \theta_i)$ for label prediction
- using regularization to prevent $h(x, \theta_i)$ from leaking the labels

Problem: regularization parameter is very sensitive

Plans and Contribution

Instead of using only $h(x, \theta)$ for training local head, one could use

- additional dataset
- activations from additional model
- activations from the pre-trained model obtained on the previous iterations of fine-tuning

We plan to explore the applicability of these dataset extensions to local head training!

References

Anonymous Authors
Label Privacy in Split Learning for Large Models with Parameter-Efficient Training
Not published, under the review

Lei Yu, Meng Han, Yiming Li, Changting Lin, Yao Zhang, Mingyang Zhang, Haiqin Weng, Yuseok Jeon, Ka-Ho Chow, Stacy Patterson.
A Survey of Privacy Threats and Defense in Vertical Federated Learning: From Model Life Cycle Perspective
arXiv:2402.03688v1