Label Privacy in Vertical Federated Learning

Kseniya Shastakova
Scientific adviser - Aleksandr Beznosikov

Moscow Institute of Physics and Technology

shestakova.ko@phystech.edu

April 23, 2024

Table of contents

High-level problem statement

We study minimization problem

$$\min_x f(x)$$

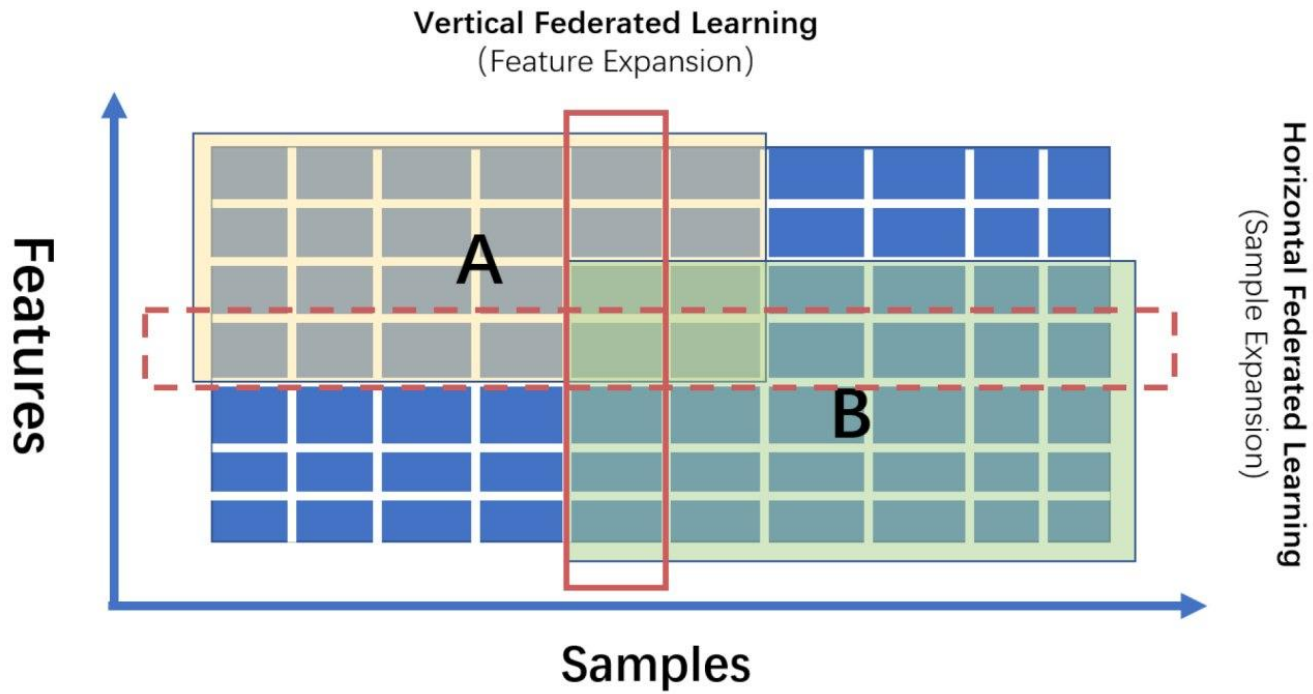
where f is usually a loss function and x is a model parameter

Introduction to VFL

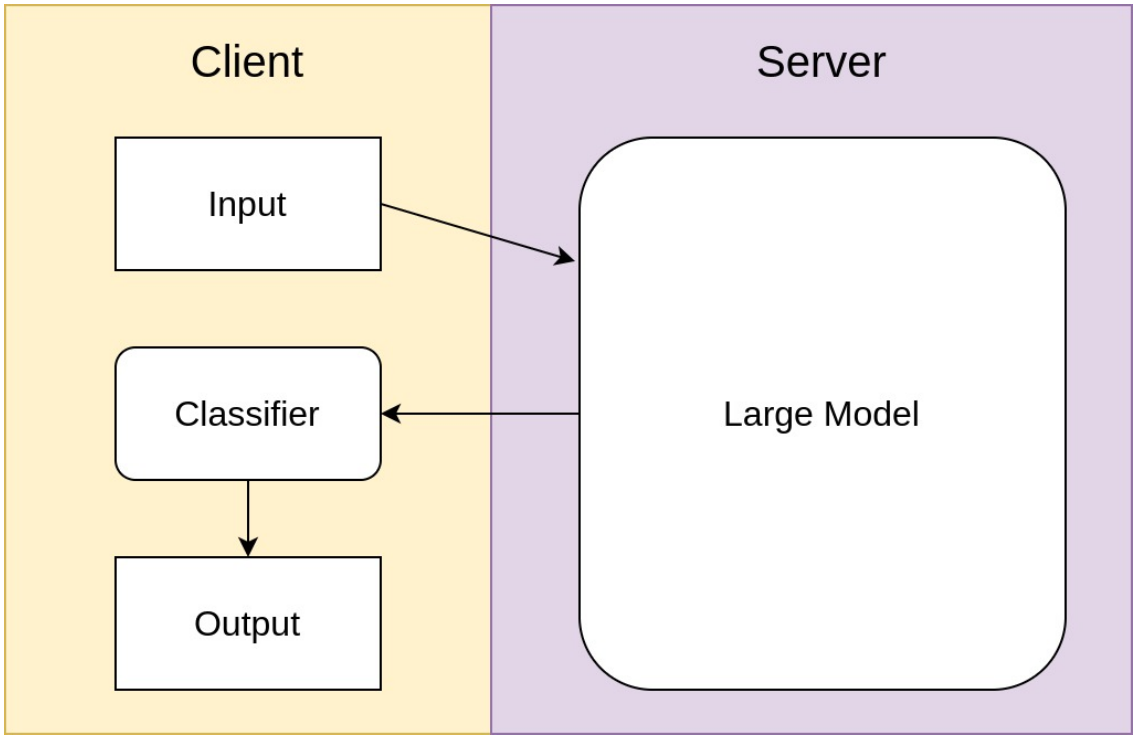
Federated Learning (FL) - several parties (a server and clients) collaboratively train a ML model

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

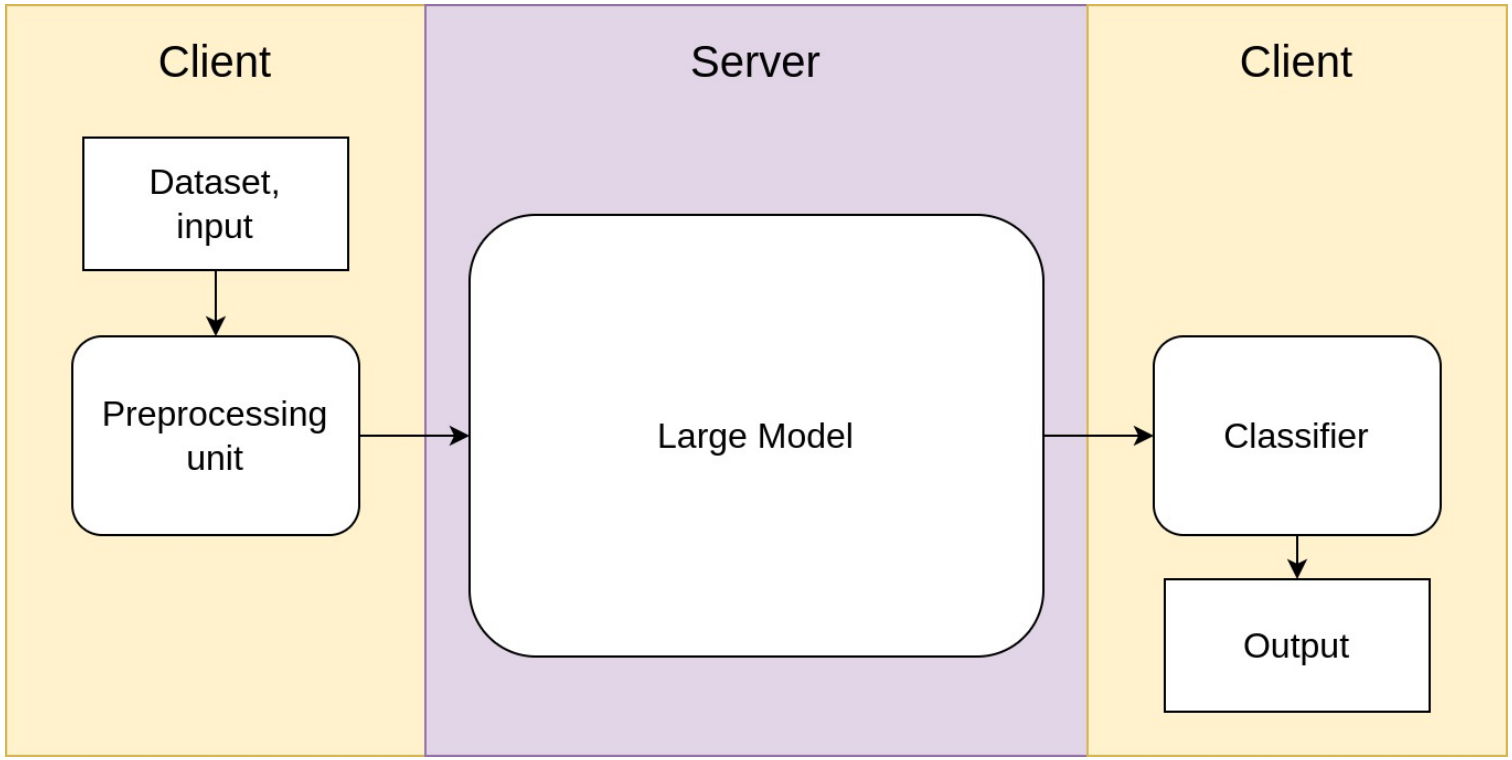
Vertical Federated Learning (VFL) - different clients have different features describing same samples



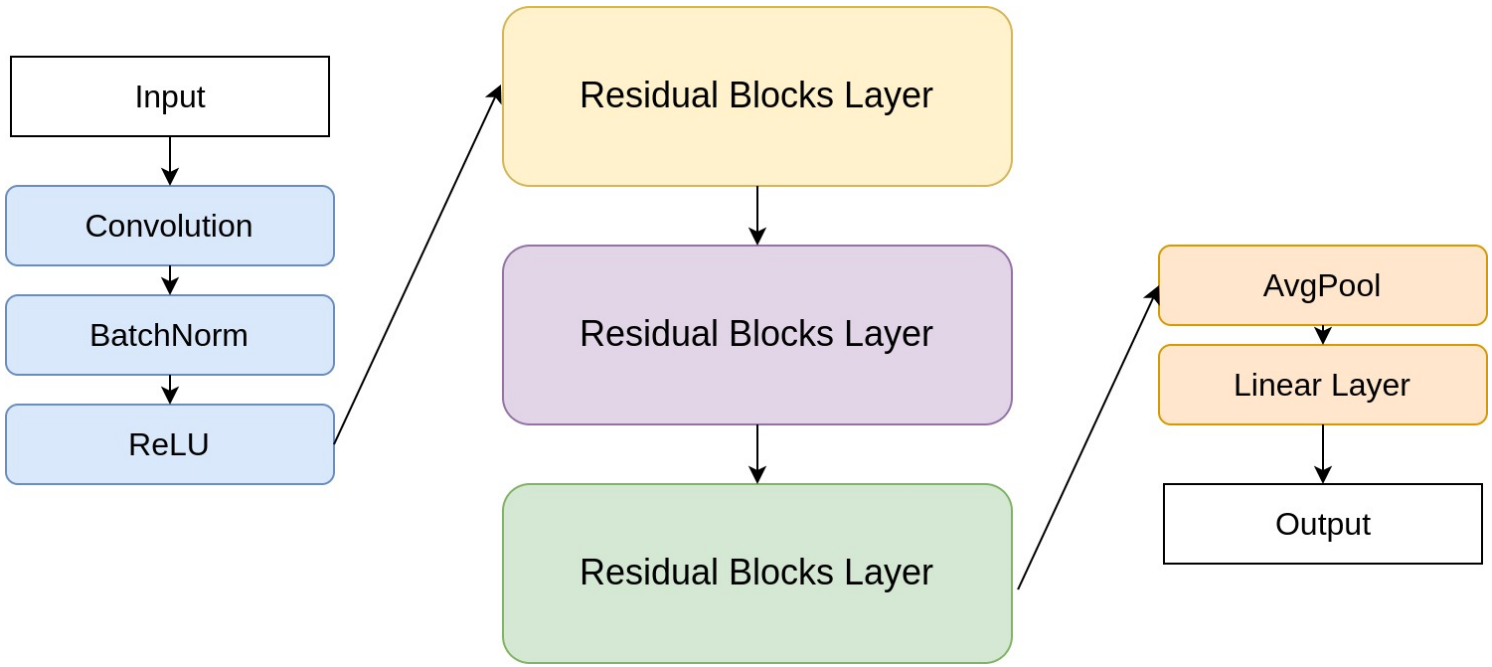
FL components in previous case



FL components in current case



ResNet


Input
Convolution
BatchNorm
ReLU
Residual Blocks Layer
Residual Blocks Layer
Residual Blocks Layer
AvgPool
Linear Layer
Output

More changes

- **Data**: binary classification on numerical data changed to picture classification (current dataset - CIFAR10)
- **Model**: server model changed from Deberta to ResNet
- **Motivation**: current work is inspired by the problem of detecting health trends based on ECG

Motivation

- Large ML models like ResNet are getting more and more popular
- CV models are in demand in medical services, but not all organizations can have the whole model on their local machines
- When using a huge model via server API, one still wants to keep one's data private

Related work

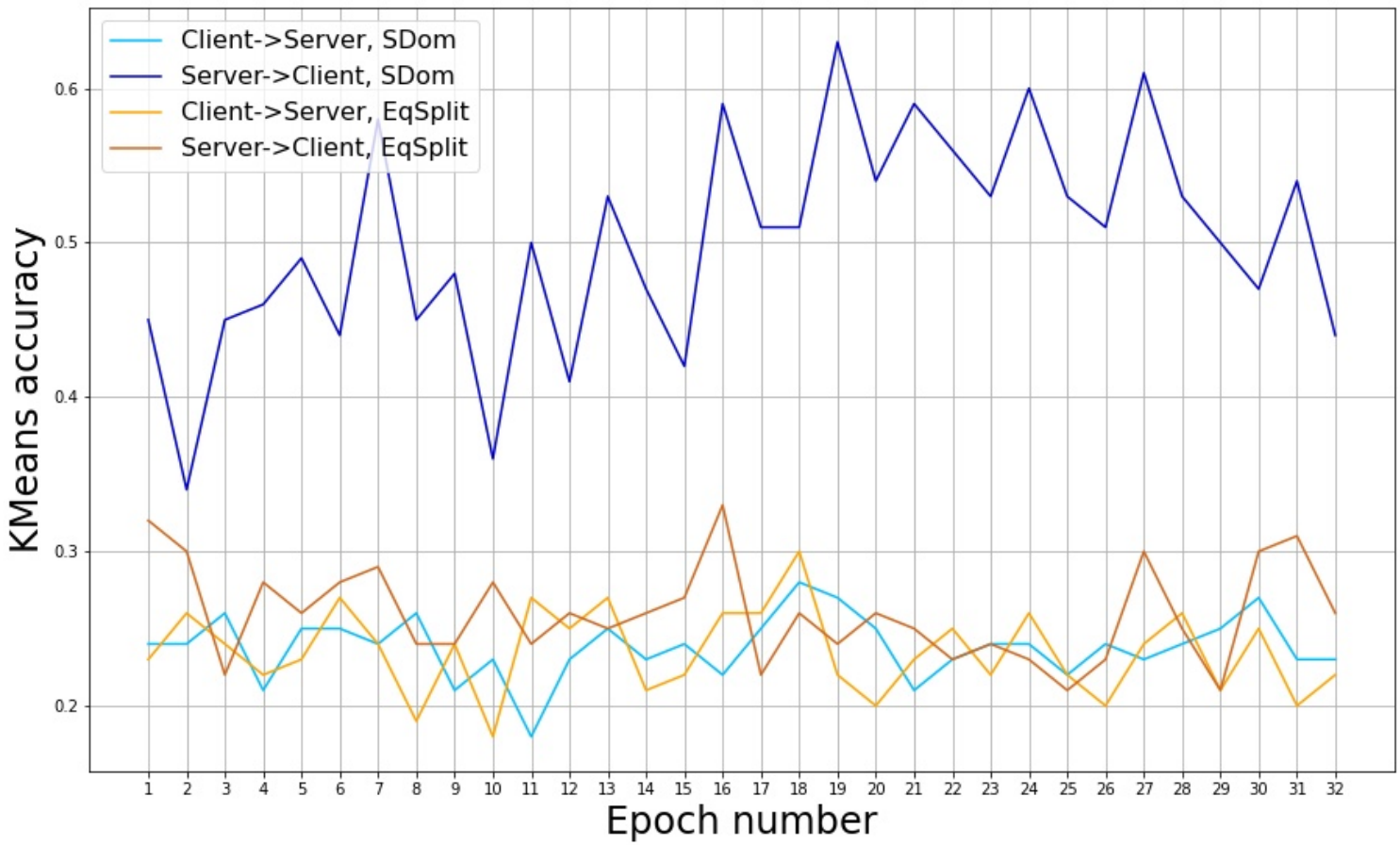
Out research aims to expand the results from [1], that shows that activations and gradients can be used to predict labels with k-means algorithm and describes possible defense based on:

- 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

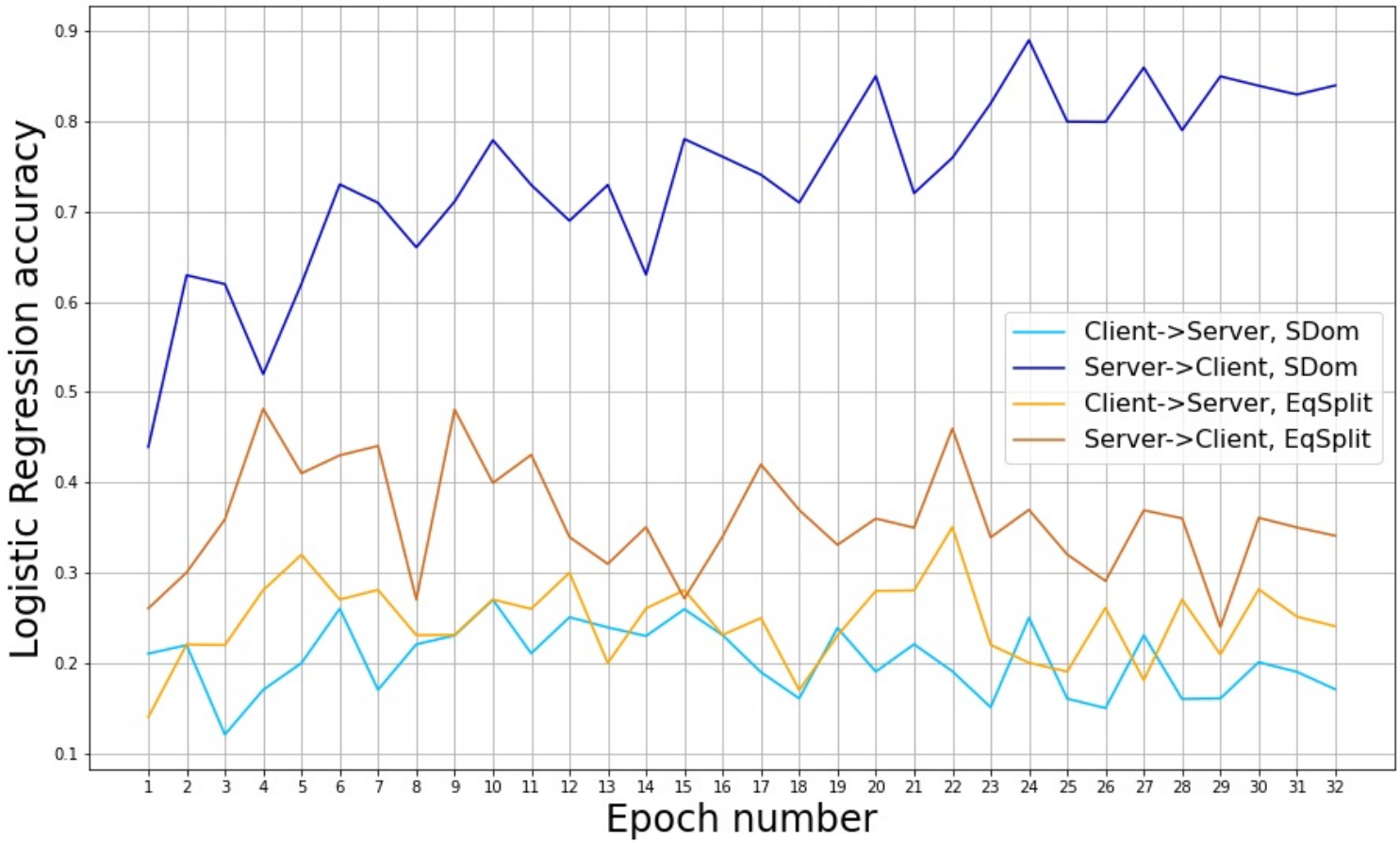
Results

- Implemented FL model with ResNet [2]
  - **Equal Splitting (EqSplit)**: each FL participant holds one Residual Blocks Layer
  - **Server Dominance (SDom)**: server holds three Residual Blocks Layers, client holds start and end blocks
- Implemented attacks on the models based on KMeans algorithm and Logistic Regression
- Started defence implementation based on regularization

KMeans Attack



Logistic Regression Attack



Plans and Contribution

- Elaborate defense strategy to repel KMeans and Logistic Regression Attacks
- Investigate more sophisticated attacks, including gradient-based methods [3]
- Implement defense strategies for successful attacks from aforementioned list

References

Anonymous Authors

Label Privacy in Split Learning for Large Models with Parameter-Efficient Training

Not published, under the review

Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun

Deep Residual Learning for Image Recognition

arxiv:1512.03385

Sanjay Kariyappa, Moinuddin K Qureshi

ExPLoit: Extracting Private Labels in Split Learning

arXiv:2402.2112.01299

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