Label Privacy in Vertical Federated Learning

Kseniya Shastakova
Scientific adviser - Aleksandr Beznosikov

Moscow Institute of Physics and Technology

shestakova.ko@phystech.edu

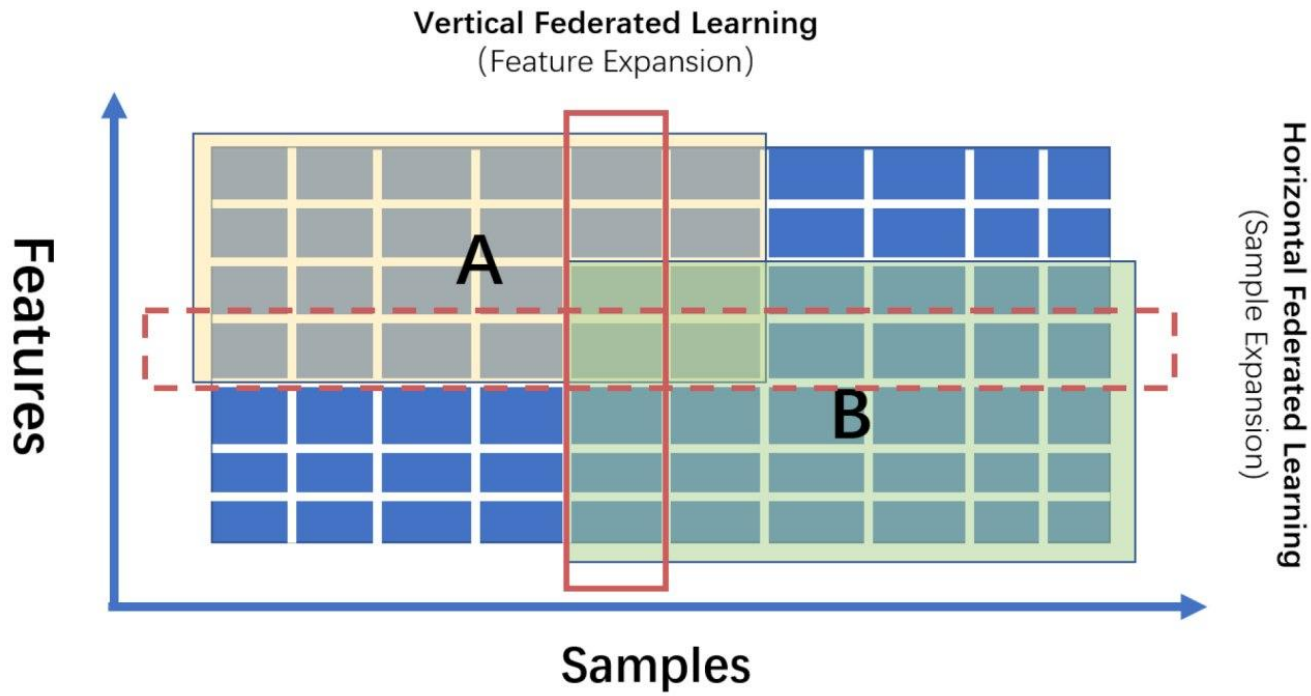
May 17, 2024

Table of contents

Introduction to VFL

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

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



Problem Statement

- Explore privacy threats for VFL systems
- Develop defense strategy to ensure data privacy

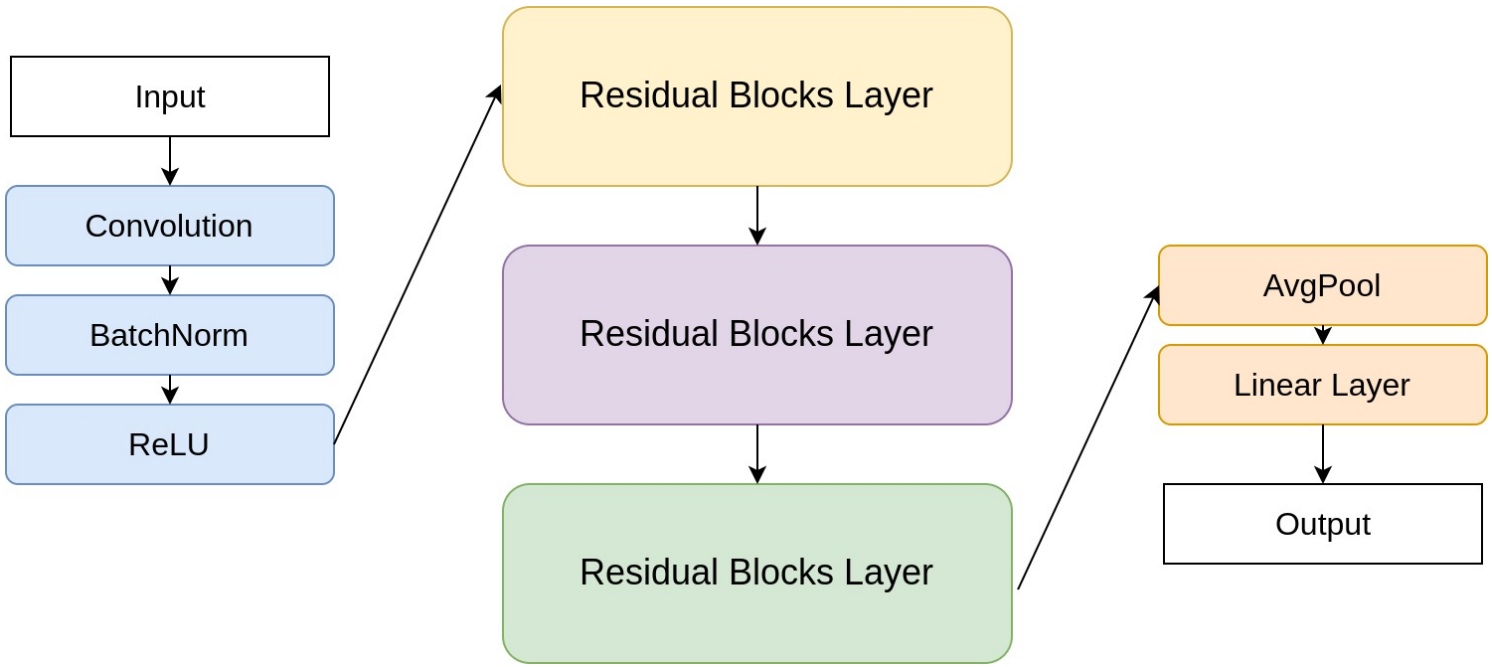
Motivation

- 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 [3]
- Current work is inspired by the problem of detecting health trends based on ECG

Stack

- **Data**: multiclass classification (current dataset - CIFAR10)
- **Model**: ResNet implemented with PyTorch library
- **Computing resources**: MIPT-Opt cluster

ResNet


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

Related work

Out research aims to expand the results from [1], that shows that activations $h(x, \theta)$ and gradients g_h 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 trained 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

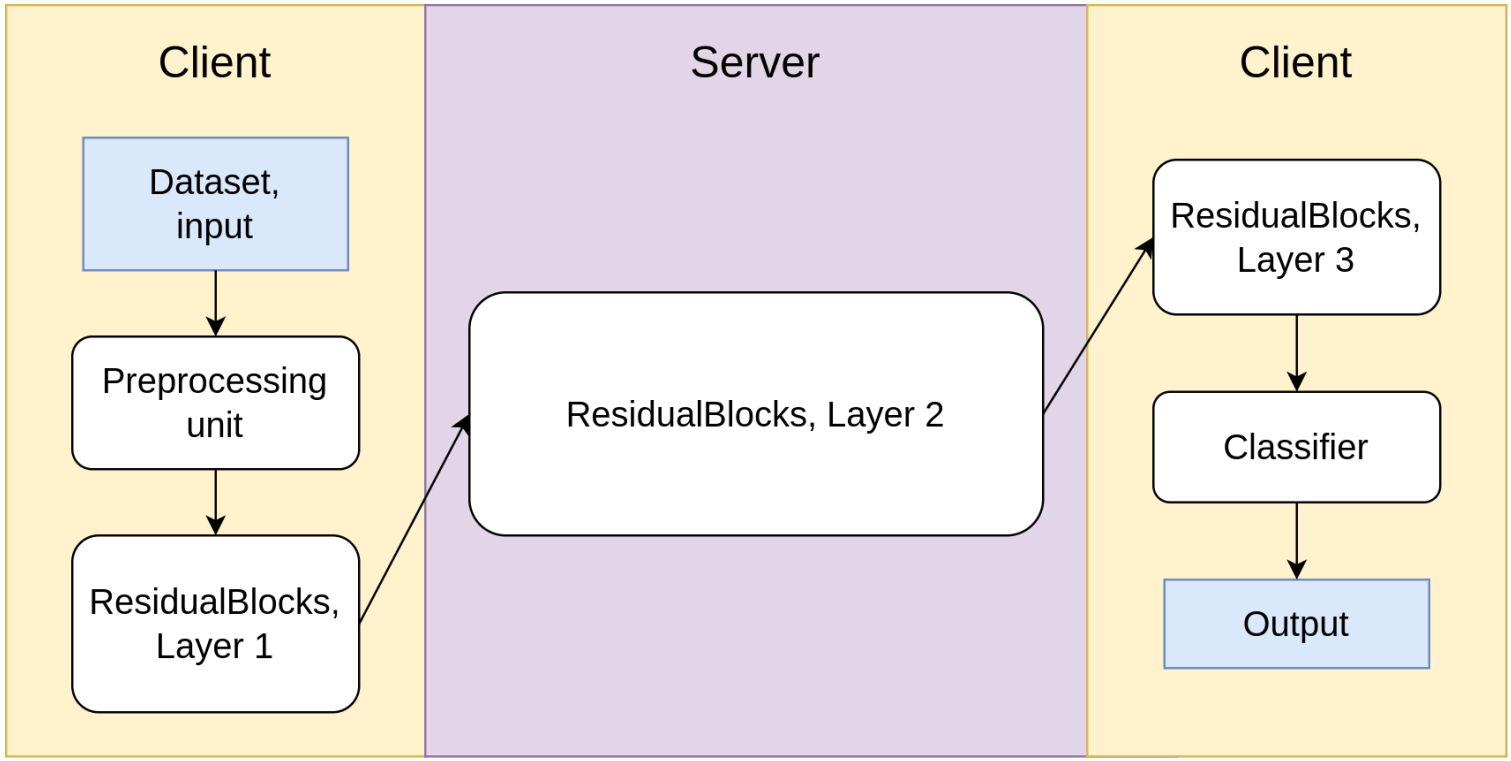
Results

- Implemented FL model for different ResNet [2] distributed architectures
- Implemented attacks on the models based on KMeans algorithm and Logistic Regression
- Implemented defence based on regularization
- Implemented defence based on additional client layers

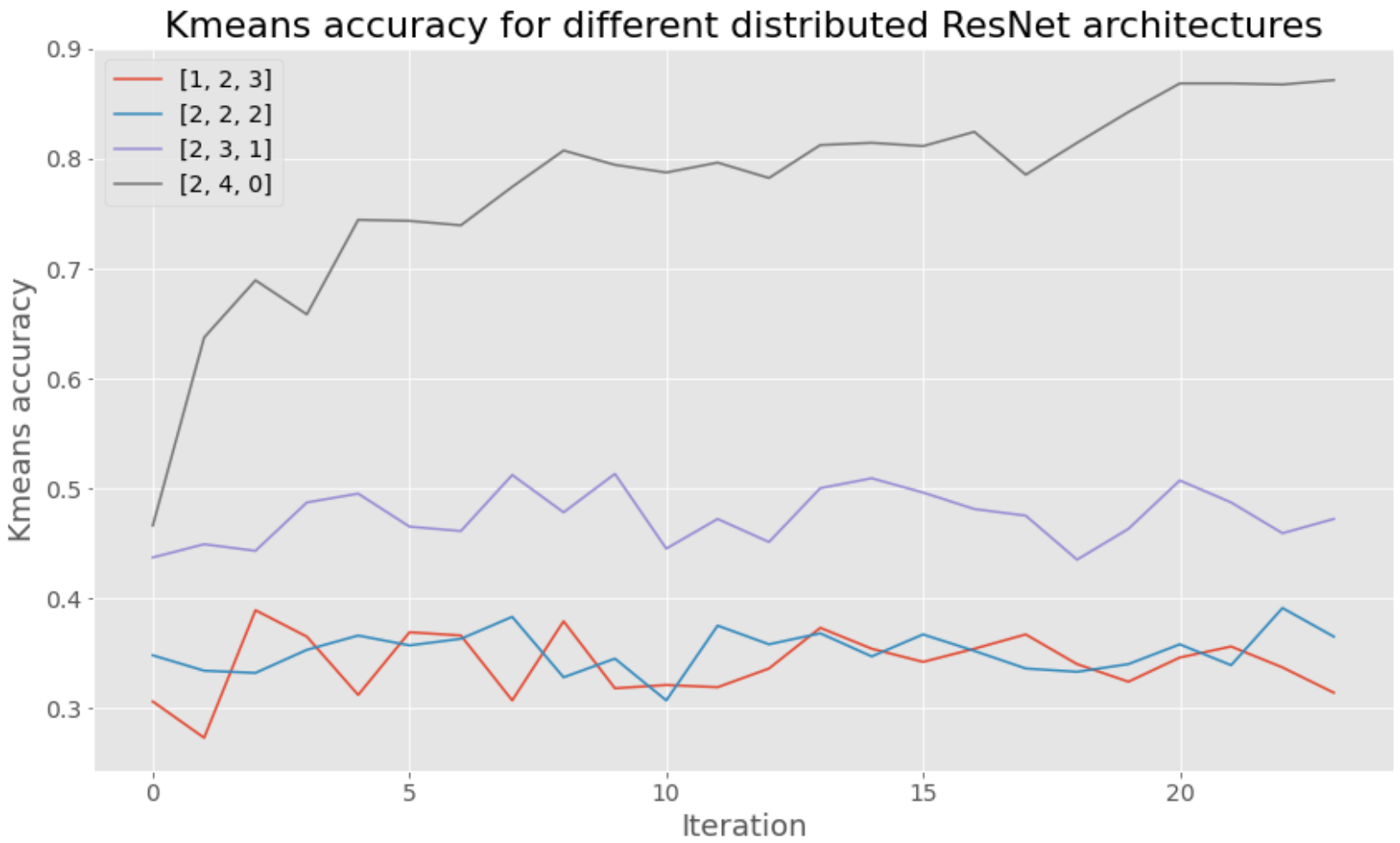
Attack Model and Defence Strategies

- **Participants**: Start Client, Server, End Client
- **Attacks**: prediction based on intermediate activations given from server to the end client
  - **Kmeans**: unsupervised clusterization method
  - **Logistic Regression**: supervised method
- **Defence**:
  - **Regularization**: loss is regularized based on LogisticRegression accuracy
  - **Additional Layers**: increase distance between server activations and final prediction

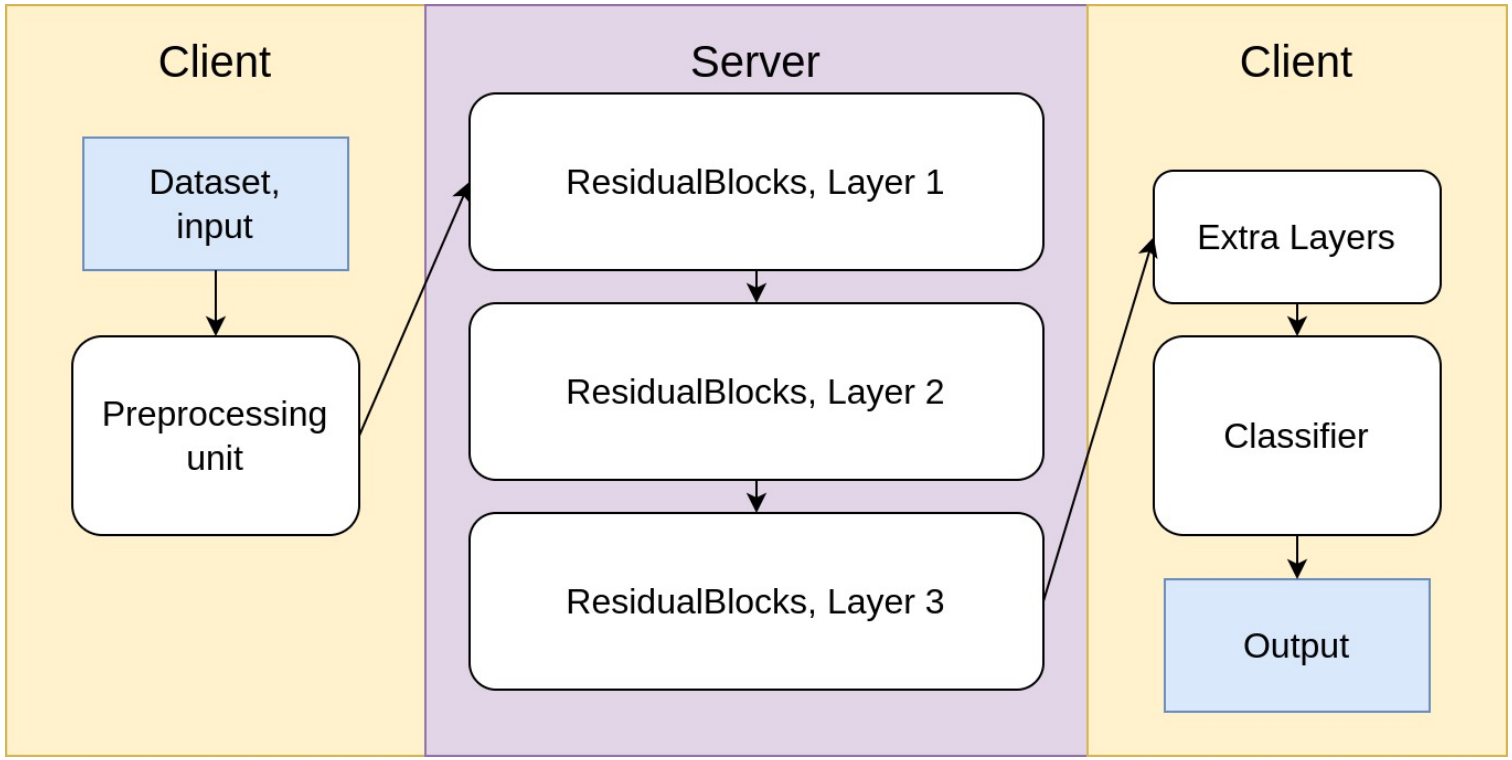
Distributed ResNet architecture



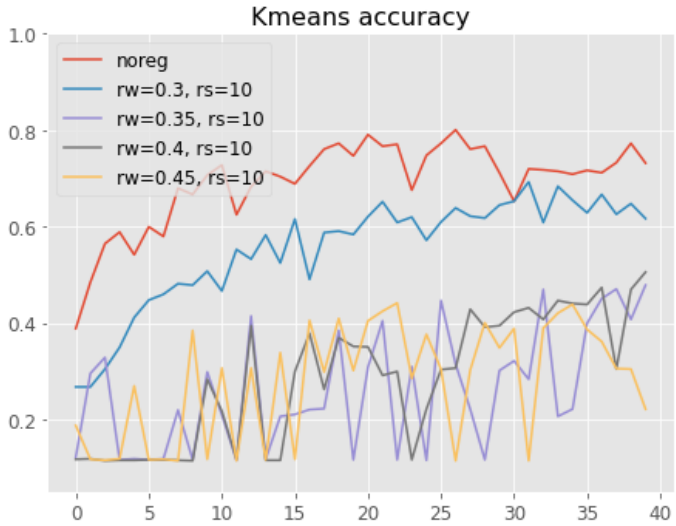
Kmeans attack accuracy

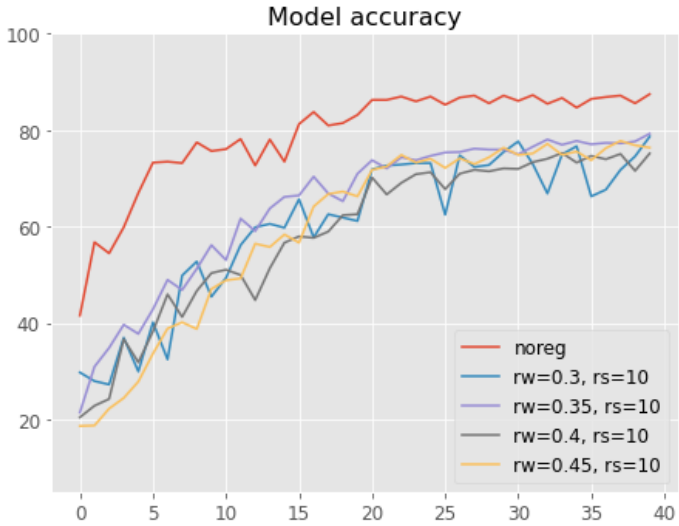


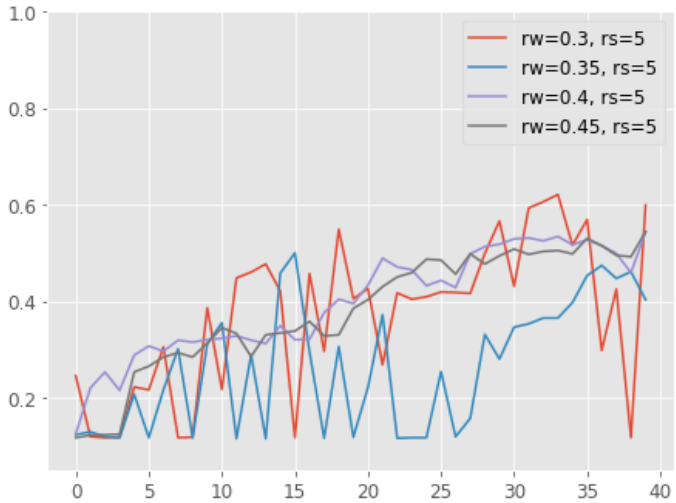
Distributed ResNet architecture

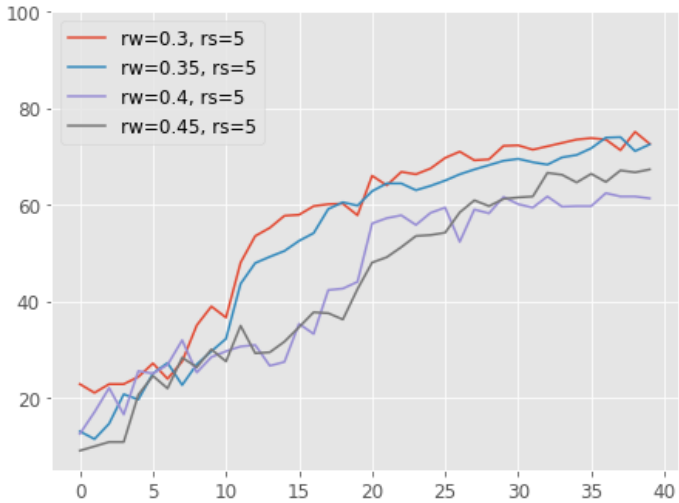


Four additional linear layers

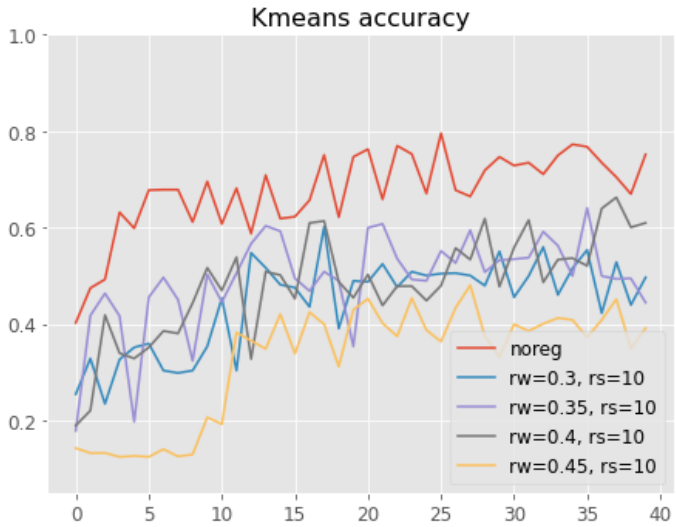


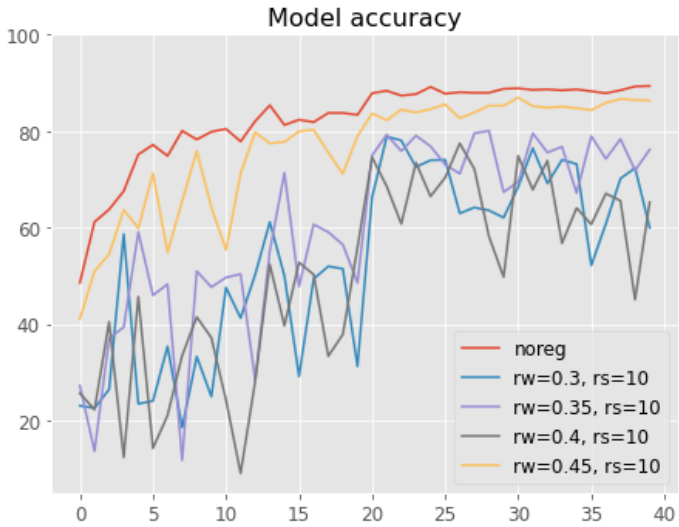


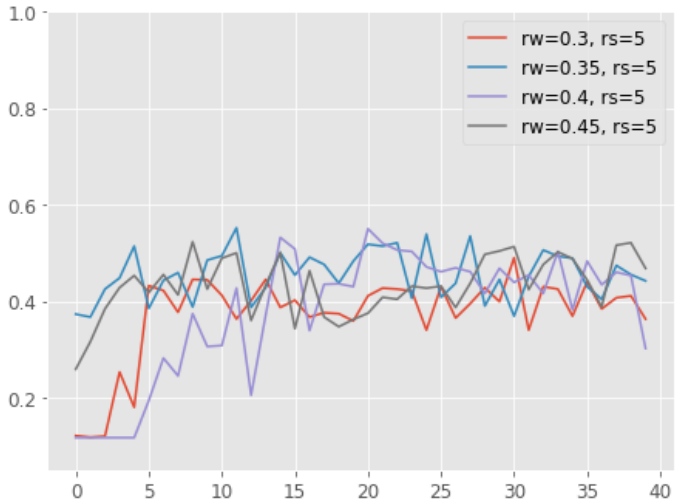




Additional ResidualBlock







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

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