Communication-Efficient Vertical Federated Learning (Extended Abstract)

Afsana Khan*, Marijn ten Thij, Anna Wilbik
Department of Advanced Computing Sciences, Maastricht University
*Correspondence: a.khan@maastrichtuniversity.nl

Federated learning (FL) is a promising approach that allows a network of autonomous organizations facing the same machine learning task to collaboratively train a global model that provides better predictive performance for all participants without requiring sensitive data sharing [1]. FL is classified into different scenarios based on the data partitioning among clients/organizations, i.e., horizontally or vertically. Horizontal federated learning (HFL), also known as homogeneous FL, refers to the scenario where clients have data with the same features but differ in the number of samples in their data. On the other hand, vertical federated learning (VFL) is used in scenarios where the clients possess different features of the same samples of data. Even though FL has addressed the issue of collaboration without compromising privacy, the repetitive updating of models during training creates significant communication overhead. Most research on communication-efficient FL [2–6] has primarily focused on the issue of large communication rounds or bandwidth in the HFL environment. However, adequate solutions are still lacking in vertically federated settings. In recent work [7], we introduced a communication-efficient vertical federated approach which uses a feature extraction technique to compress local data of clients. The compressed data from the clients are then aggregated to train the final machine learning model. As a result, clients collaborate by sharing compressed (latent) representations of their raw data without jeopardizing their privacy and security. In addition, the entire process is limited to a single communication round.

In a vertical federated setting, clients possessing relevant disjoint data are interested in training a global machine learning model without exposing their raw data to each other. One of the clients is assumed to have labeled data (guest party) and the rest (host parties) have unlabeled data. The objective of the guest party/client is to be able to use the data from the host parties/clients in order to perform better predictions for incoming new data, without compromising the privacy of data for itself as well as for other clients. The proposed method, as shown in Figure 1, is based on feature extraction techniques that reduce the dimensionality of data by removing redundancy. The feature extraction methods generally obtain new generated features by combining and transforming the original feature set, thus giving it a new latent representation. The first step of the proposed method begins with performing feature extraction, also referred to as feature compression, on the local data of each client to generate latent representations of the local data. To perform feature extraction, we experimented using two techniques; Principle Component Analysis (PCA) and Undercomplete Autoencoders (AE). In the latter step, the compressed local data of the clients are aggregated to train the global model. The latent data differ significantly from the original data of the clients but still contain relevant important information. As a result, sharing latent data from clients poses no risk of raw data exposure and also improves performance.

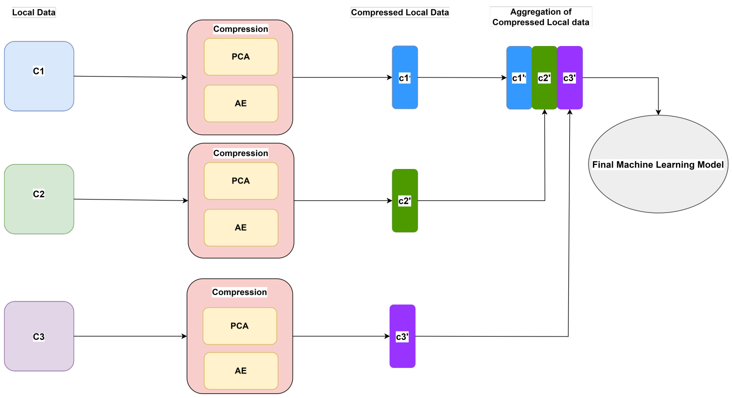

Fig. 1: Architecture of Proposed Method

For performing experiments several datasets were chosen based on their public availability and the number of samples & features varying from small to large, so that the robustness of the proposed method could be evaluated. The chosen datasets were as follows:

- Adult Income Dataset [8]
- Heart Disease Dataset [9]
- Wine Quality Dataset [10]
- Rice MSC Dataset [11]

The performance metrics (Accuracy and F1-Score) obtained through the proposed method were compared against the centralized learning system and also the guest client's learning system using only its local data. The experimental results using the proposed method with PCA as a feature extraction technique showed that the aggregated model outperformed the local model of the guest client across all datasets. In case, when autoencoders were used as the feature extraction technique, similar results were obtained. However, in case of datasets with large dimension, the autoencoders became lossy. The high losses of the autoencoders indicated that the local data were not effectively compressed containing most of the original information. Hence, a drop in the performance of the aggregated model was observed. However, it is clear that if the autoencoder losses are minimized, the proposed method will perform well, just as it did in case of other datasets. By adding more hidden layers and properly tuning the hyperparameters during training, the performance of the autoencoders can be enhanced. The experimental setup and elaborate analysis of the results is available in our original paper [7].

References

1. Qiang Yang, Yang Liu, Yong Cheng, Yan Kang, Tianjian Chen, and Han Yu. Federated learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 13(3):1–207, 2019.
2. Mingzhe Chen, Nir Shlezinger, H Vincent Poor, Yonina C Eldar, and Shuguang Cui. Communication-efficient federated learning. *Proceedings of the National Academy of Sciences*, 118(17):e2024789118, 2021.
3. Neel Guha, Ameet Talwalkar, and Virginia Smith. One-shot federated learning. *arXiv preprint arXiv:1902.11175*, 2019.
4. Anirudh Kasturi, Anish Reddy Ellore, and Chittaranjan Hota. Fusion learning: A one shot federated learning. In *International Conference on Computational Science*, pages 424–436. Springer, 2020.
5. Qinbin Li, Bingsheng He, and Dawn Song. Practical one-shot federated learning for cross-silo setting. *arXiv preprint arXiv:2010.01017*, 2020.
6. Felix Sattler, Simon Wiedemann, Klaus-Robert Müller, and Wojciech Samek. Robust and communication-efficient federated learning from non-iid data. *IEEE transactions on neural networks and learning systems*, 31(9):3400–3413, 2019.
7. Afsana Khan, Marijn ten Thij, and Anna Wilbik. Communication-efficient vertical federated learning. *Algorithms*, 15(8):273, 2022.
8. Dheeru Dua and Casey Graff. Adult. UCI Machine Learning Repository, 2017.
9. Robert Detrano, Andras Janosi, Walter Steinbrunn, Matthias Pfisterer, Johann-Jakob Schmid, Sarbjit Sandhu, Kern H Guppy, Stella Lee, and Victor Froelicher. International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American journal of cardiology*, 64(5):304–310, 1989.
10. Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision support systems*, 47(4):547–553, 2009.
11. Ilkay Cinar and Murat Koklu. Classification of rice varieties using artificial intelligence methods. *International Journal of Intelligent Systems and Applications in Engineering*, 7(3):188–194, 2019.