SparseFedPP: sparse federated learning for hardware-constrained edge-devices

Adriano Guastella[1], **Alessio Mora**[1], **Paolo Bellavista**[1],
Lorenzo Sani[2], **Alexandru-Andrei Iacob**[2], **Nicholas D. Lane**[2]
[1]University of Bologna, Italy
[2]University of Cambridge, UK

Federated Learning (FL) enables learning predictive models without centralizing raw data, marking a paradigm shift toward privacy-preserving machine learning [1]. It enables effective collaboration by alternating between on-device training, communication, and aggregation of locally fine-tuned models. However, FL also introduces several challenges: (a) dealing with communication bottlenecks and (b) execution under computational constraints on edge devices. Using sparse models is a promising solution to address such concerns since suppressing less relevant model parameters permits the reduction of both the communication costs and the computational costs of the process, especially when specialized hardware can exploit sparse operations [2]. However, to be effective during training, the sparsification procedure must account for the iterative nature of federated optimization and the potentially heterogeneous data held on learners (i.e., clients). For example, if each client were to naively and independently sparsify their parameters after training, the resulting sparsification masks may differ if their data heterogeneity induces dissimilar models.

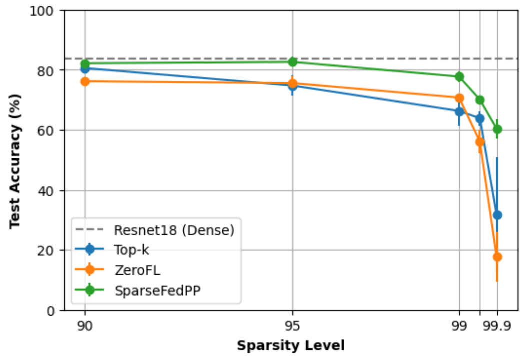

Figure 1: Test Accuracy of Different Sparsification Methods on ResNet-18 on CIFAR10.

This work introduces *SparseFedPP*, a training accelerator inspired by *Powerpropagation* [3] and *SWAT* [4], to improve the efficiency of cross-device FL. Through the integration of Powerpropagation which induces clients towards naturally highly sparse models, we have effectively addressed some primary challenges associated with sparse networks in FL. Firstly, faster convergence during training of the global model is achieved, which leads to better consensus on the sparsity mask among all clients, resulting in improved performance. The achieved sparsity in the model is then utilized to analyze the model's layer sensitivity, allowing for accurate sparsification of the activation during the backward pass, following an approach similar to the one proposed in SWAT.

With SparseFedPP, we were able to apply a high sparsity ratio to the model with minimal or no performance loss, achieving **sparsity levels up to 99.9%**. Thanks to a more aware sparse distribution among the layers, we outperformed *ZeroFL*, a previous integration of SWAT in FL [5]. Fig. 1 shows the difference in performance degradation between SparseFedPP, ZeroFL, and a naive Top-k sparsification applied after local training [6], particularly at high levels of sparsity. Furthermore, when coupled with a lossless model compression technique such as Compressed Sparse Row (CSR) [7], *SparseFedPP* produces a remarkable **145x speed-up in communication** costs.

Experimental results have demonstrated a significant reduction in computational operations, leading to a potential speed-up in on-device inference and training time. Additionally, there is a significant decrease in memory consumption during training, both on-device, facilitated by the sparsification of activations saved for the backward pass, and in communication, due to the substantial reduction in model dimensions. The decrease in communication bandwidth usage also accelerates the training rounds. These achievements make FL more scalable and efficient, enabling less capable devices to train models in FL even in constrained networks.

References

[1] Brendan McMahan et al. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[2] Zihao Zhao et al. Towards efficient communications in federated learning: A contemporary survey. *Journal of the Franklin Institute*, 360(12):8669–8703, 2023.

[3] Jonathan Schwarz et al. Powerpropagation: A sparsity inducing weight reparameterisation. In *35th Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[4] Md Aamir Raihan and Tor M Aamodt. Sparse weight activation training. In *Advances in Neural Information Processing Systems*, December 2020.

[5] Xinchi Qiu et al. Zerofl: Efficient on-device training for federated learning with local sparsity, 2022.

[6] Shaohuai Shi et al. Understanding top-k sparsification in distributed deep learning, 2019.

[7] W.F. Tinney and J.W. Walker. Direct solutions of sparse network equations by optimally ordered triangular factorization. *IEEE, 55(11):1801–1809, 1967. doi: 10.1109/PROC.1967.6011*, 1967.