ZeroFL: Efficient On-device Training for Federated Learning with Local Sparsity

Xinchi Qiu [1*], Javier Fernandez-Marques[2*], Pedro P.B. Gusmao[1], Yan Gao[1],
Titouan Parcollet[3], Nicholas D. Lane[1]
[1] Department of Computer Science and Technology, University of Cambridge
[2]Department of Computer Science, University of Oxford
[3]Laboratoire Informatique d'Avignon,Avignon Universite

In order to adjust the memory and compute footprints of complex ML models to the FL setting, the research community has presented a number of approaches including such as federated dropout [1] or different aggregation strategies that enable faster convergence [2]. Other optimization techniques such as quantization and sparsity have been used in the context of FL but mostly as a way to reduce communication costs [3–5] but not to accelerate on-device training.

The use of sparse operations at training time has recently been shown to be an effective technique to accelerate training in centralised settings [6]. The resulting models are as good or close to their densely-trained counterparts despite reducing by up to 90% their FLOPs budget and, resulting in an overall up to 3.3× training speedup. Acceleration is achieved by performing sparse convolutions during the forward and/or backward pass, which requires at least one of the operands (i.e. inputs, weights, gradients) to be sufficiently sparse and, software and hardware support for such operations. However, it is unclear how the different FL-specific challenges (i.e. data imbalance, stateless clients, periodic aggregation) will restrict the quality of the global model.

This work considers the challenges and opportunities of inducing high levels of sparsity to accelerate training on-device for FL workloads, and provides the following contributions: (1) A study on the unique aspects that arise when introducing sparsity at training time; (2) We then propose ZeroFL, a framework that relies on highly sparse operations to accelerate on-device training. Models trained with ZeroFL and 95% sparsity achieve up to 2.3% higher accuracy compared to competitive baselines obtained from adapting a state-of-the-art sparse training framework to the FL setting. (3) a comprehensive analysis on CIFAR-10, FEMNIST and SpeechCommands datasets in terms of model performance and communication costs.

References

[1] S. Horvath, S. Laskaridis, M. Almeida, I. Leontiadis, S. I. Venieris, and N. D. Lane, "Fjord: Fair and accurate federated learning under heterogeneous targets with ordered dropout," *arXiv preprint arXiv:2102.13451*, 2021.

[2] S. J. Reddi, Z. Charles, M. Zaheer, Z. Garrett, K. Rush, J. Konečný, S. Kumar, and H. B. McMahan, "Adaptive federated optimization," in *ICLR*, 2021.

[3] L. Liu, J. Zhang, S. Song, and K. B. Letaief, "Hierarchical quantized federated learning: Convergence analysis and system design," 2021.

[4] M. M. Amiri, D. Gunduz, S. R. Kulkarni, and H. V. Poor, "Federated learning with quantized global model updates," 2020.

[5] O. Shahid, S. Pouriyeh, R. M. Parizi, Q. Z. Sheng, G. Srivastava, and L. Zhao, "Communication efficiency in federated learning: Achievements and challenges," 2021.

[6] M. A. Raihan and T. M. Aamodt, "Sparse weight activation training," *arXiv preprint arXiv:2001.01969*, 2020.