Federated LLM Training with Heterogeneous Mobile Clients

Andrzej Szablewski, Lorenzo Sani, Nicholas D. Lane

May 2025

Federated learning is a method for training machine learning models across multiple, usually weakly connected clients in a distributed setting – ultimately, these would be transformative if such clients can be mobile and embedded devices. Since the emergence of foundation models and their rapidly growing sizes [1], FL can be considered a communication-efficient alternative to standard SGD-like optimisers [2]. At the same time, the pre-training of state-of-the-art Large Language Models (LLMs) is mainly dominated by large corporations due to the requirement of significant computing resources, making it prohibitively expensive for individuals or communities to train their own large models. To tackle this challenge, FL may allow large-scale collaborative efforts towards pre-training on commodity PC hardware, mobile devices, and the edge. Furthermore, such projects may involve participants' high-quality private data, which is usually unavailable in public data repositories. However, this collaborative federated training often involves parties with diverse specifications, including heterogeneous hardware, varying compute resources, or network bandwidths.

With this research, we demonstrate the results of the study on the choice of compute-optimal hyperparameters. While several works explored this problem in the general context of federated learning [3, 4, 5, 6], we focus on pre-training transformer architectures in a single-epoch regime on varying devices, including those of a mobile, wearable and embedded variety. We build on the results indicating the possibility of using larger learning rates and smaller batch sizes, compared to the standard, centralised training methodologies [7]. These discoveries are particularly promising in the context of training large models on mobile devices with limited resources. Notably, combined with the federated learning scenario, the heterogeneous environment results in multiple local (client-level) and global (federation-level) hyperparameters, increasing the search space for optimal configuration. We mainly focus on the impact of the local and global batch sizes, optimiser configurations, and their schedules, as well as the number of federated rounds and steps in each of them. We compare the approaches based on factors such as the total compute, training wall-time, and energy usage. Our study concludes with a guide to the selection of optimal federated hyperparameters to make such choices handy in complex scenarios such as cross-device federated learning. We anticipate such information being of huge value to the mobile community as they begin to build various kinds of these systems (viz. networks of robots, IoT devices and wearables).

Furthermore in our work, we will show additional attempts at lowering the compute required for LLM training, involving the disjoint training of suitable neural architecture elements. One such element is the multi-head attention module of a transformer's block, which consists of independent attention heads. Selectively training only some of the heads on each federated client further reduces the total compute and communication bandwidth needed for collaborative model training.

References

[1] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models, 2022.

[2] H. Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data, 2023.

[3] Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data, 2020.

[4] Blake Woodworth, Kumar Kshitij Patel, Sebastian U. Stich, Zhen Dai, Brian Bullins, H. Brendan McMahan, Ohad Shamir, and Nathan Srebro. Is local sgd better than minibatch sgd?, 2020.

[5] Yi Zhou, Parikshit Ram, Theodoros Salonidis, Nathalie Baracaldo, Horst Samulowitz, and Heiko Ludwig. Single-shot hyper-parameter optimization for federated learning: A general algorithm analysis, 2022.

[6] Mikhail Khodak, Renbo Tu, Tian Li, Liam Li, Maria-Florina Balcan, Virginia Smith, and Ameet Talwalkar. Federated hyperparameter tuning: Challenges, baselines, and connections to weight-sharing, 2021.

[7] Lorenzo Sani, Alex Iacob, Zeyu Cao, Royson Lee, Bill Marino, Yan Gao, Dongqi Cai, Zexi Li, Wanru Zhao, Xinchi Qiu, and Nicholas D. Lane. Photon: Federated llm pre-training, 2024.