PhD School: Efficient Optimization in Resource-Constrained Federated Learning

Ouiame Marnissi
PhD student at College of Computing, Mohammed VI Polytechnic University (UM6P)
Benguerir, Morocco
ouiame.marnissi@um6p.ma

Abstract

Federated learning is a distributed machine learning scheme that enables collaborative model training without compromising data privacy. However, implementing federated learning in real wireless networks comes with some challenges as high dimensional updates are frequently exchanged over resource-constrained networks. In my work as a PhD student, I investigate resource optimization for efficient federated learning. In particular, my objective is to minimize the communication overhead (time and energy) while ensuring a good convergence, in resource-constrained and heterogeneous environments.

CCS Concepts

• **Computing methodologies** → **Federated learning**.

Keywords

Federated Learning, Client Sampling, Compression, Resource allocation, Decentralized Federated Learning.

1 Research Activities

Federated learning (FL) has emerged as a promising technique that enables multiple devices to collaboratively learn a global model without sharing their personal data [6]. Therefore, as opposed to classical distributed machine learning (ML) techniques -where devices need to share their data with a centralized server-, not only the clients' privacy is preserved, but also the computation load is shared at the edge network. However, implementing FL in real wireless networks comes with several key challenges. In particular, devices are usually powered by limited batteries and communicate with limited resources. Therefore, transmitting high-dimensional ML updates generates a considerable cost in terms of learning time and energy consumption. Furthermore, clients in FL usually hold unbalanced and non-IID data and have different communication and computation capabilities (e.g., bandwidth, CPU frequency). This heterogeneity of the clients (including statistical and system heterogeneity) combined with the limited wireless resources incur a significant energy consumption and a large latency which is impractical in real-life situations.

To address these challenges *client sampling* techniques have been investigated in the literature, attempting to answer the question: *how can we select the optimal subset of devices in order to accelerate the convergence while respecting the wireless resources?* [1]. Selecting the best devices reduces the number of communication rounds needed for convergence. However, the size of FL parameters shared in each round is still very high.

To reduce the communication costs, *compression* techniques (including sparsification and quantization) are proposed [7]. Sparsification aims to send a sparse vector of weights containing only selected elements of the vector, whether quantization consists of reducing the number of digits that are used to encode the gradient vector. This leads us to a second important research question: *how to devise an FL approach that chooses the best sparsification and quantization level at each round in order to respect the wireless resources while meeting the target model accuracy?*

In classical FL, training occurs under the coordination of a central server that aggregates the coordinates from all the participating devices. Consequently, the server can become a bottleneck, especially with an increasing number of connected devices. Moreover, a single server is prone to failure and attacks which compromise the entire learning process. As an alternative, *decentralized FL* (DFL) overcomes these limitations by enabling direct data exchanges between the nodes [2]. Therefore, the network is no longer constrained by a single server, which allows better scalability and eliminates the risk of single-point failure. We carefully reviewed the literature related to DFL, and one important question arose: *how to efficiently optimize wireless resources in DFL with a time-varying network topology?*

After identifying the various research activities we aim to work on, we outline the research methods in the next section.

2 Research methods

To efficiently conduct our research, we first thoroughly review the literature, specifically the recent advances focusing on resource optimization through device selection, and compression techniques in both classical and decentralized FL.

- We study and understand the existing device sampling strategies in the current state-of-the-art (e.g., uniform, weighted, statistical, and importance sampling)
- We review the different existing compression techniques (e.,g., sparsification and quantization)
- We analyze how the existent device sampling and compression techniques impact model convergence and communication resources.

Second, we identify key research objectives that we intend to focus on. In particular, we try to answer the question: *after identifying the research activities and reviewing the related state-of-the-art, how can we contribute to advancing the field?* In the following are some of the answers we came up with.

- Develop and evaluate client sampling methods to accelerate model convergence while respecting the wireless resource constraints with a special focus on the non-IID case.

EWSN, December 10–13, 2024, Abu Dhabi, UAE

- Propose an efficient FL algorithm that optimizes compression levels to reduce communication overhead (i.e., time and energy) while ensuring convergence in a heterogeneous framework.
- Design a communication-efficient DFL algorithm that optimizes performance in time-varying network topologies. The proposed architecture should integrate compression techniques to reduce communication overhead as each device only needs to send a compressed model to its selected peers. Furthermore, and most importantly, this approach should consider the network fluctuation in each round and select the optimal subset of peers for an efficient energy minimization.

In the next section, we discuss the progress we have made so far, including the paper we have published, the results obtained, and our ongoing works.

3 Results and Ongoing works

3.1 Published Results

- **Optimal Client Sampling:** In [4], we propose an efficient selection technique whereby the subset of participating devices is determined based on the norm of their gradients. Extensive simulations are drawn to demonstrate the performance of our approach compared to the benchmarks. In [5], we propose FedHSC, a framework that considers both system and statistical heterogeneity. Specifically, at each communication round, the clients are sampled based on their data properties combined with the importance of their local learning update. After completing their local training, the selected clients share compressed updates with the server for aggregation. The compression rate is adjusted for each client to meet the communication delay requirement. In fact, compression techniques have been widely used to reduce the communication overhead in FL. In the following, we study the effect of an optimal compression (in both computation and communication) in considerably saving wireless resources.
- **Optimal Compression:** In [3], we propose a compression framework that tackles energy efficiency. In particular, our method selects the optimal quantization parameters in both training and communication, and the best sparsification levels in transmission to minimize the total energy consumed while respecting the time constraint and ensuring convergence. To achieve this, we formulate an optimization problem that we efficiently solve with low time complexity. Our simulation results confirm the outperformance of our method compared to the state of the art.

3.2 Ongoing works

Currently, we are investigating resource optimization in DFL. Our aim is to leverage our acquired knowledge in device selection and compression within FL to apply and extend these concepts in the context of DFL. After extensive analysis, we were able to establish a relationship between the convergence bound, the compression levels and the set of connected peers. We are working on an efficient optimization algorithm that aims to learn these unknown parameters with the objective of energy minimization. Once the parameters are solved, we plan various simulations to validate our approach.

4 Challenges

As a PhD student, I am currently facing several technical and non-technical challenges that are impacting the progress of my research. While I am constantly trying to solve them, I am reaching out for feedback and advice from experts and fellow researchers, hoping their insights can help me overcome these obstacles.

4.1 Technical Challenges

During my research, I have encountered several technical obstacles that have slowed my progress. Below, I have listed some of these key challenges.

- Data heterogeneity: The quality of model training degrades with highly skewed and non-IID data, thus impacting the convergence and the model performance.
- Compression optimization: Compression is known to mitigate the communication overhead in FL. However, too much compression impedes the model convergence. Thus, it is difficult to balance the compression levels, model accuracy, and communication performance.
- Time-varying topologies in DFL: Difficulties in reaching consensus and global model convergence in a fluctuating device participation framework.

4.2 Non-Technical Challenges

In addition to the aforementioned technical challenges, several non-technical challenges have impacted the smooth progress of my research. I have listed the most important ones in the following.

- Publication pressure: Managing the pressure to publish in prestigious and top-tier venues.
- Self-depreciation and lack of motivation: Dealing with self-doubt, in particular when facing paper rejections or when my work is not progressing as I would like.
- Networking: As a Moroccan, I face difficulties in attending high-level conferences and events. Furthermore, lack of experience in networking makes it hard to engage with experts and renowned researchers.
- Time management: Balancing research activities and commitments with personal life, especially with heavy workload and tight deadlines.

5 Conclusion

Thanks to its privacy-preserving property, FL is applied in different contexts (e.g., healthcare, agriculture, autonomous driving). As a PhD student, I still have a lot to learn, both in terms of technical expertise and personal development. Hence, I need guidance and support from experts to help me navigate this adventure and successfully tackle my research journey. My wish is for my research to contribute to advancing the efficiency of FL algorithms and, in doing so, to be an active part of the AI revolution.

References

[1] L. Fu, H. Zhang, Ge Gao, M. Zhang, and X. Liu. 2023. Client selection in federated learning: Principles, challenges, and opportunities. *IEEE Internet of Things Journal*

(2023).
[2] W. Liu, L. Chen, and W. Zhang. 2022. Decentralized federated learning: Balancing communication and computing costs. *IEEE Transactions on Signal and Information Processing over Networks* 8 (2022), 131–143.
[3] O. Marnissi, H. El Hammouti, and E. H. Bergou. 2024. Adaptive sparsification and quantization for enhanced energy efficiency in federated learning. *IEEE Open Journal of the Communications Society* (2024).
[4] O. Marnissi, H. El Hammouti, and E. H. Bergou. 2024. Client selection in federated learning based on gradients importance. In *AIP Conference Proceedings*, Vol. 3034. AIP Publishing.
[5] O. Marnissi, H. EL Hammouti, and E. H. Bergou. 2024. Efficient Client Sampling with Compression in Heterogeneous Federated Learning. In *IEEE INFOCOM 2024-IEEE Conference on Computer Communications Workshops*. IEEE, 1–2.
[6] B. McMahan, E. Moore, D. Ramage, S. Hampson, and Blaise A. y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*. PMLR, 1273–1282.
[7] S. U Stich, J. Cordonnier, and M. Jaggi. 2018. Sparsified SGD with memory. *Advances in neural information processing systems* 31 (2018).