A Review on Research Trends of Optimization for Client Selection in Federated Learning

Jaemin Kim*, Chihyun Song*, Jeongyeup Paek*, Jung-Hyok Kwon†, and Sungrae Cho*
*School of Computer Science and Engineering, Chung-Ang University, Seoul 06974, South Korea.
†Hallym University Smart Computing Laboratory, Hallym University, Chuncheon 24252, South Korea.
Email:{jmkim, chsong}@uclab.re.kr, 02superjh@gmail.com, {jpaek, srcho}@cau.ac.kr

***Abstract*—Federated Learning (FL) presents a paradigm shift in data-driven model training, allowing for the collaborative learning of a shared model while keeping all the training data on the client side, hence avoiding central data accumulation. Within this framework, the selection of clients becomes a pivotal element impacting learning efficiency, model performance, and communication overhead. Engaging all clients simultaneously in the learning process can lead to inefficiencies, particularly when considering resource constraints and diverse data distributions. Consequently, there has been a burgeoning interest in research dedicated to developing optimization techniques tailored for strategic client selection to address these challenges. This paper delves into the various optimization strategies proposed in the realm of FL for user selection, scrutinizing their objectives, methodologies, and outcomes. [1] Our review indicates that these techniques not only contribute to alleviating bandwidth constraints and reducing computational loads but also enhance model performance by intelligently choosing clients that provide the most informative updates. Furthermore, we discuss the trade-offs involved in such optimization processes, like the balance between model accuracy and training time, and highlight potential paths for future research that may pave the way for more sophisticated and efficient federated systems.**

***Index Terms*—federated learning, client selection, optimization, reinforcement learning.**

I. Introduction

As we advance into an era where data is ubiquitously generated at an unprecedented scale, the imperative for sophisticated data analysis paradigms that can process this deluge while respecting user privacy has never been more critical. This paper delves into the burgeoning field of Federated Learning (FL), a paradigm shift in machine learning that allows for the collaborative training of algorithms across multiple decentralized devices or clients, all while keeping the data localized.

Federated Learning posits a significant advantage in preserving the privacy and security of data, as it eliminates the necessity for data to be transmitted to or stored in a central location. This aspect is especially pertinent in scenarios where data sensitivity or regulatory compliance, such as GDPR in Europe, is of paramount importance. However, while FL promises a multitude of benefits, it also introduces complex challenges that must be surmounted to realize its full potential. Among these, the strategy for selecting which clients participate in the learning process stands out as a critical determinant of the efficiency and effectiveness of the federated model. [2]

Client selection in FL is far from trivial—this procedure requires careful consideration of several factors, including data distribution, resource availability, and network reliability, among others. The heterogeneity of clients in terms of their data and computational capabilities introduces additional layers of complexity. Some clients may have data that is more relevant or of higher quality for the learning task, while others may be limited by their computational resources or network connectivity.

An effective client selection mechanism is vital for several reasons: to enhance the learning performance, to ensure the swift convergence of the model, and to manage the communication overhead, which can be substantial in a distributed environment. Moreover, client selection impacts the fairness of the learning process, as it determines which data points contribute to the model's evolution, potentially influencing the model's bias and variance.

The focus of this paper is to present a comprehensive review of the optimization techniques for client selection that have been proposed in recent FL research. We aim to assess the state-of-the-art methods, identify their strengths and limitations, and provide a structured overview of this crucial aspect of FL. Through this exploration, we will shed light on how these techniques can be tailored to different FL scenarios, such as cross-device and cross-silo settings, and the implications of these choices on the resulting federated learning models.

With this introduction, we set the stage for a deep dive into the dynamics of client selection within the FL framework, exploring the multifaceted approaches that aim to harmonize the trade-offs between data privacy, system efficiency, and learning performance.

II. Related Work

A. Federated Learning

Federated learning employs an iterative framework that involves recurrent communication between a central server and individual devices (clients). [3] This recurring communication is often referred to as a 'communication round,' and each communication round consists of several distinct phases, as outlined below:

Identify applicable funding agency here. If none, delete this.

- **1.Initialization Phase**
  Before the commencement of the first round, the central server initializes the global model weight as ω_0.
- **2. Client Selection Phase**
  At the beginning of each round, a user fraction parameter, denoted as C, is set. The central server then selects users according to the specified fraction, and subsequently, it dispatches the current global model weight to each selected user's device.
- **3. Update Phase**
  In this phase, each user conducts computations based on the global model and their local dataset. Subsequently, they transmit the updated local model weight back to the central server.
- **4. Aggregation Phase**
  The central server amalgamates the received updated local model weights into the global model.
- **5. Termination Phase**
  If the global model reaches convergence with a certain loss threshold, the process terminates. Otherwise, it proceeds back to phase 2 for another round of communication and updates.

B. FedAvg

The FedAvg algorithm provides a way to train a model across multiple devices or clients and average them to update the entire model. This promotes a distributed learning approach where the model is trained while keeping the data local on the client device itself, as opposed to training centrally on a central server, while taking into account the security and privacy of personal data.

The main idea of FedAvg is:

- **1. Local model updates**
  Each client device updates its model using local data.
- **2. Central server integration**
  A central server aggregates model updates learned from each client to create a full model.
- **3. Average model distribution**
  The central server sends the average model to the client devices, which is used as the local model for the next round of training

[4]

III. Client Selection in Federated Learning

The cornerstone of federated learning (FL) lies in its decentralized approach, allowing a myriad of devices and entities to partake in a collective learning process while ensuring the individual's data remains local. This paradigm shift caters to the ever-growing need for privacy preservation and computational efficiency in today's data-centric society. However, the cornerstone of this emerging field is the strategy of client selection, a nuanced and critical task that demands meticulous attention.

Certainly! Here's a more concise version of the provided text in English:

In federated learning, participant equality varies with device diversity, from smartphones to IoT devices, each contributing uniquely to the learning algorithm. This diversity, while enriching, poses challenges in selecting the right client subset to ensure effective learning and broad data representation. Therefore, client choice is crucial, directly affecting model quality and utility.

The focus has shifted from mere data aggregation to assessing data quality. High-quality data from clients contributes more significantly to a robust learning model than larger amounts of lower-quality data. The key challenge is identifying and prioritizing data that most benefits the learning process, considering the dynamic nature of data to maintain model relevance under varying conditions.

Diversity in client data, a double-edged sword, enhances model generalization but risks incorporating outliers or noisy data that could skew results. Achieving balance requires mechanisms to distinguish beneficial heterogeneity from harmful anomalies. [5] The integrity of federated learning depends heavily on client data reliability. Poor or unstable data can cause significant learning trajectory deviations, leading to ineffective models. [6] Thus, mechanisms to assess and ensure data quality and stability before integration into the learning process are essential.

Client heterogeneity affects communication efficiency and computational costs in federated learning. Selecting clients who can quickly transmit data and have adequate computational power is key to improving the efficiency of the learning cycle. Developing advanced algorithms to assess various factors and make trade-offs between data quality, client reliability, and computational efficiency is essential for effective client selection and overall success in federated learning.

IV. State-of-the-Art on Optimization Techniques

In seeking to enhance the performance and efficiency of Federated Learning (FL), a pivotal aspect revolves around the optimization of client selection and the mitigation of data-related challenges such as non-IID distribution and privacy concerns. Recent advancements have brought forward innovative methodologies that intertwine machine learning algorithms, reinforcement learning, and dimensionality reduction techniques to address these complexities.

A. Deep Reinforcement Learning for Device Selection

The selection of client devices for training in FL has been reframed as a Deep Reinforcement Learning (DRL) challenge. The sophisticated nature of device ecosystems, encompassing heterogeneity in data, computational capabilities, and availability, necessitates an adaptive approach that DRL caters to. By treating each round of FL as a Markov Decision Process (MDP), we have an environment where the state encompasses the global model weights in conjunction with each client's model weights. A DRL agent, leveraging a Double Deep Q-learning Network (DDQN), is trained to select a subset of clients for local training. The reward mechanism is contingent upon the accuracy of the global model, ascertained through a validation set, driving the agent towards the expedited attainment of target performance metrics. [7]

B. Principal Component Analysis in Federated Learning

In the proposed DRL-based selection process, the privacy-preserving ethos of FL is upheld—no actual data samples are required from the clients, only their model weights are essential. This is critical in maintaining user privacy. To further streamline this process, dimensionality reduction techniques such as Principal Component Analysis (PCA) can be utilized. PCA aids in condensing the model weight information, thus expediting the communication process while preserving the critical information necessary for effective device selection. This dimensionality reduction does not obfuscate the divergence between local model weights, which is vital in informing the selection process by the DRL agent. [8] [9] [10]

C. XGBoost Combined with Federated Learning

An exemplary integration of advanced machine learning algorithms with FL is the application of eXtreme Gradient Boosting (XGBoost). XGBoost, an enhanced iteration of the Gradient Boosting Decision Tree (GBDT), harnesses second-order gradient information to provide more rapid and accurate convergence, along with regularization terms that diminish overfitting risks. The synergy of XGBoost and FL is predicated on XGBoost's adaptability to distributed machine learning frameworks, enabling it to function seamlessly within FL's paradigm. In the context of our horizontal FL deployment, the process commences with the server disseminating a pre-trained model to the participants. Subsequently, each node computes and encrypts the model parameters locally before sending them back to the server for aggregation. The updated model, after server-side aggregation, is redistributed to the nodes for decryption and local updating. This iterative process persists until the performance of the model reaches the desired threshold or the maximum number of iterations is achieved. [11]

In summation, these innovative approaches—XGBoost's robustness, DRL's adaptability, and PCA's efficiency—coalesce to form a formidable strategy in overcoming the prevailing challenges within Federated Learning. By meticulously addressing the issues of client heterogeneity, dynamic availability, and data privacy, this strategy seeks to usher in a new era of efficiency and efficacy for Federated Learning frameworks. [12]

V. Conclusion

This paper highlights the significance of client selection in Federated Learning (FL), emphasizing its impact on system efficacy. We explored various strategies, including XGBoost integration, Deep Reinforcement Learning (DRL) for dynamic client selection, and model weight dimensionality reduction. These methods are crucial for enhancing network efficiency, accuracy, and addressing non-IID data challenges in FL.

However, the complexity of data diversity and communication constraints in real-world scenarios presents challenges in client selection algorithm design. Addressing these while maintaining privacy and security is essential for future development. Future research should focus on evaluating and refining client selection algorithms in FL, particularly in real-time and large-scale contexts [13] [14]

Acknowledgment

This work was supported in part by the National Research Foundation of Korea (NRF), South Korea grant funded by the Korea government (MSIT) (No. 2022R1A4A5034130), in part by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2024-RS-2022-00156353) supervised by the IITP (Institute for Information and Communications Technology Planning and Evaluation), South Korea.

References

[1] N. Pavlidis, V. Perifanis, T. P. Chatzinikolaou, G. C. Sirakoulis, and P. S. Efraimidis, "Intelligent client selection for federated learning using cellular automata," *arXiv preprint arXiv:2310.00627*, 2023.

[2] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[3] C. Smestad and J. Li, "A systematic literature review on client selection in federated learning," *arXiv preprint arXiv:2306.04862*, 2023.

[4] S. Mayhoub and T. M. Shami, "A review of client selection methods in federated learning," *Archives of Computational Methods in Engineering*, pp. 1–24, 2023.

[5] M. Aiche, S. Ouchani, and H. Bouarfa, "Leader-assisted client selection for federated learning in iot via the cooperation of nearby devices," in *International Conference on Machine Learning for Networking*. Springer, 2022, pp. 169–177.

[6] Y. Shi, Z. Liu, Z. Shi, and H. Yu, "Fairness-aware client selection for federated learning," in *2023 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2023, pp. 324–329.

[7] H. Wang, Z. Kaplan, D. Niu, and B. Li, "Optimizing federated learning on non-iid data with reinforcement learning," in *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*. IEEE, 2020, pp. 1698–1707.

[8] M. Tang, X. Ning, Y. Wang, J. Sun, Y. Wang, H. Li, and Y. Chen, "Fedcor: Correlation-based active client selection strategy for heterogeneous federated learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 102–10 111.

[9] S. Chen, O. Tavallaie, M. H. Hambali, S. M. Zandavi, H. Haddadi, S. Guo, and A. Y. Zomaya, "Boosting client selection of federated learning under device and data heterogeneity," *arXiv preprint arXiv:2310.08147*, 2023.

[10] P. Singhal, S. R. Pandey, and P. Popovski, "Greedy shapley client selection for communication-efficient federated learning," *arXiv preprint arXiv:2312.09108*, 2023.

[11] P. Liu, T. Gao, and C. Li, "Optimization of federated learning communications with heterogeneous quantization," in *2022 IEEE 22nd International Conference on Communication Technology (ICCT)*. IEEE, 2022, pp. 292–296.

[12] T. Nishio and R. Yonetani, "Client selection for federated learning with heterogeneous resources in mobile edge," in *ICC 2019-2019 IEEE international conference on communications (ICC)*. IEEE, 2019, pp. 1–7.

[13] L. Best, E. Foo, H. Tian, and Z. Jadidi, "Client selection frameworks within federated machine learning: The current paradigm," in *Emerging Smart Technologies for Critical Infrastructure*. Springer, 2023, pp. 61–83.

[14] A. Gouissem, Z. Chkirbene, and R. Hamila, "A comprehensive survey on client selections in federated learning," *arXiv preprint arXiv:2311.06801*, 2023.