TOWARDS EFFICIENT AND PRACTICAL FEDERATED LEARNING

Ahmed M. Abdelmoniem [1]

ABSTRACT

Federated learning (FL) is increasingly becoming the norm for training models over distributed and private datasets. Major service providers rely on leveraging end-user data for training global ML models to improve services such as text auto-completion, virtual keyboards, and item recommendations.

In this poster, we describe our efforts which are motivated by the urging need for exploring the promising prospects and imminent challenges towards the facilitation of the adoption of federated learning for service providers. The prospects are motivated by the growing interest and momentum towards the adoption of privacy preservation and 5G/6G technologies. Whereas, the challenges that hinder wide FL adoption are mainly the resource and user heterogeneity, and communication overhead. One major challenge is the system heterogeneity which can hinder the progress or convergence of the FL-trained models. Specifically, FL-trained models in practice require a significant amount of time (days or even weeks) because FL tasks execute in highly heterogeneous environments where devices only have widespread yet limited computing capabilities and network connectivity conditions.

Therefore, we present our initial efforts focused on studying training FL models in heterogeneous environments. We note that heterogeneity can have a detrimental impact on both the model quality and fairness. Following our analytical and empirical study, we present our efforts to design efficient yet practical mitigation methods to limit the impact of system heterogeneity. Our results in various settings and benchmarks show that, compared to state-of-the-art methods, the proposed approaches can significantly improve the quality and fairness of the FL-trained models while reducing the resource consumption on target devices.

1 INTRODUCTION

Recent growing interest for Edge AI deployments accompanied by the increasing privacy concerns has introduced notable changes in the training methods of customer-facing machine learning (ML) models. Many big data analytic rely on developing and training forecasting, voice recognition, or image classification models to serve most end-user applications (Abdelmoniem & Canini, 2021b; Abdelmoniem et al., 2021a). In the new training method, the model is shipped to the user's device and training is conducted on the private user data locally on the end devices moving towards the right direction from training on centrally-collected data which poses privacy concerns on the transfer of private data. The abundance of rich and timely datasets on end-user devices motivates the computations to be outsourced to a distributed set of end-devices at the edge.

Federated Learning (FL) is a recent paradigm that has sparked great research interest to enable learning collaboratively over distributed datasets (Konečný et al., 2016; McMahan et al., 2017; Bonawitz et al., 2019). In FL, with the help of a central aggregation server, the end devices train a global model on their private data without transferring the private data over the network (Nasr et al., 2019). Even though, the paradigm was initially proposed by the industry (e.g., (Hard et al., 2018; Yang et al., 2018; Bonawitz et al., 2019)) to solve a practical problem, it has also seen tremendous interest from academia (Caldas et al., 2018; Yang et al., 2021; Mohri et al., 2019; Lai et al., 2021; Nasr et al., 2019)).

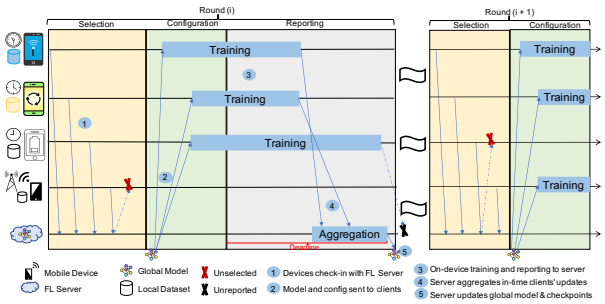

Figure 1: Phases of FL in heterogeneous environment

Brief Intro to FL Training Process: We briefly explain the most commonly used architecture for cross-community (or device) federated learning as depicted in Figure 1. Typically, N clients owning end-devices storing their own pri-

[1]Queen Mary University of London, UK. Correspondence to: Ahmed M. Abdelmoniem <ahmed.sayed@qmul.ac.uk>.

vate datasets which have a common features or structure. The devices collaborate to learn a global model via a centralized aggregation FL server which controls the progress:

1. The devices check-in with the FL server, and then the server typically selects a sample of devices for training and pushes a copy of the up-to-date global model.
2. The devices perform an equal number of local optimization steps as determined by task designer.
3. The FL server performs secure aggregation of the local models pushed by the clients.
4. The FL server updates the state of the global model.

Heterogeneity Challenges: Numerous studies have shown that FL faces major challenges hindering its wide adoption in practice (Yang et al., 2018; Bonawitz et al., 2019; Yang et al., 2019). Due to the distributed nature of the training on a large number of heterogeneous clients (in terms of data distributions, computational and communication capabilities, and/or availability), tackling the heterogeneity is considered one of the grand challenges. The following summarizes the main sources of heterogeneity in FL:

- **Data Heterogeneity:** mainly because the data distribution on end-devices is non-identical independent distributed (Non-IID). Therefore, data popularity and observation bias are introduced into the model. For instance, it is intuitive that some clients produce more data samples or higher quality data than others (Mohri et al., 2019; Bonawitz et al., 2019).
- **Device Heterogeneity:** is caused by the variable capabilities of the end-devices. The differences in end-devices' system configuration would result in missing the reporting deadline, partial updates or failures to communicate the updates. Typically, slower devices are more susceptible to failures or missing deadlines (Bonawitz et al., 2019; Lai et al., 2021).
- **behavioral Heterogeneity:** is caused by the behavior of the end-users. For instance, the user behaviour influences the end-devices' status (e.g., idle, charging, or connected to WiFi) which results in sampling bias affecting model quality (Yang et al., 2021).

Poster Summary: In this poster, we present our preliminary steps and main findings towards studying and mitigating the effects of heterogeneity in FL environments:

1. Through extensive empirical study, we find that system heterogeneity can result in degradation of 5X on average for model quality (Abdelmoniem et al., 2022).
2. AQFL applies per-device custom model quantization to reduce the stragglers and hence improve the model quality (Abdelmoniem & Canini, 2021a).
3. We present RELAY which implements staleness-aware aggregation and intelligent participant selection algorithms to improve resource usage with minimal impact on time-to-accuracy (Abdelmoniem et al., 2021b)

REFERENCES

Abdelmoniem, A. M. and Canini, M. Towards mitigating device heterogeneity in federated learning via adaptive model quantization. In *ACM EuroMLSys*, 2021a.

Abdelmoniem, A. M. and Canini, M. DC2: Delay-aware Compression Control for Distributed Machine Learning. In *IEEE INFOCOM*, 2021b.

Abdelmoniem, A. M., Elzanaty, A., Alouini, M.-S., and Canini, M. An Efficient Statistical-based Gradient Compression Technique for Distributed Training Systems. In *MLSys*, 2021a.

Abdelmoniem, A. M., Sahu, A. N., Canini, M., and Fahmy, S. A. Resource-efficient federated learning. *arXiv 2111.01108*, 2021b.

Abdelmoniem, A. M., Ho, C.-Y., Papageorgiou, P., and Canini, M. Empirical analysis of federated learning in heterogeneous environments. In *ACM EuroMLSys*, 2022.

Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., Kiddon, C., Konečný, J., Mazzocchi, S., McMahan, H. B., Overveldt, T. V., Petrou, D., Ramage, D., and Roselander, J. Towards Federated Learning at Scale: System Design. In *MLSys*, 2019.

Caldas, S., Duddu, S. M. K., Wu, P., Li, T., Konečný, J., McMahan, H. B., Smith, V., and Talwalkar, A. Leaf: A benchmark for federated settings. *arXiv 1812.01097*, 2018.

Hard, A., Rao, K., Mathews, R., Ramaswamy, S., Beaufays, F., Augenstein, S., Eichner, H., Kiddon, C., and Ramage, D. Federated learning for mobile keyboard prediction, 2018.

Konečný, J., McMahan, H. B., Yu, F. X., Richtarik, P., Suresh, A. T., and Bacon, D. Federated Learning: Strategies for Improving Communication Efficiency. In *Workshop on Private Multi-Party Machine Learning - NeurIPS*, 2016.

Lai, F., Zhu, X., Madhyastha, H. V., and Chowdhury, M. Efficient Federated Learning via Guided Participant Selection. In *USENIX OSDI*, 2021.

McMahan, H. B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*, 2017.

Mohri, M., Sivek, G., and Suresh, A. T. Agnostic federated learning. In *ICML*, 2019.

Nasr, M., Shokri, R., and Houmansadr, A. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *IEEE Symposium on Security and Privacy (SP)*, 2019.

Yang, C., Wang, Q., Xu, M., Chen, Z., Bian, K., Liu, Y., and Liu, X. Characterizing impacts of heterogeneity in federated learning upon large-scale smartphone data. In *The Web*, 2021.

Yang, Q., Liu, Y., Chen, T., and Tong, Y. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology*, 10(2), 2019.

Yang, T., Andrew, G., Eichner, H., Sun, H., Li, W., Kong, N., Ramage, D., and Beaufays, F. Applied Federated Learning: Improving Google Keyboard Query Suggestions. *arXiv 1812.02903*, 2018.