On the Performance Gains of Federated Learning Edge Caching in Vehicular Internet of Things

Lilian C. Mutalemwa and Seokjoo Shin*
Department of Computer Engineering, Chosun University, Gwangju 61452, South Korea
Email: lilian.mutalemwa@gmail.com, *sjshin@chosun.ac.kr (corresponding author)

Abstract—**In fifth generation (5G) wireless networks, one main challenge of edge artificial intelligence is to train machine learning models by aggregating a large amount of data that are distributed at different edge devices. Sending training data to a centralized cloud server introduces prohibitive communication overhead. Moreover, many types of data contain privacy-sensitive personal data. Federated learning (FL) is a recently proposed machine learning paradigm that allows to collaboratively train a shared model for many users without direct access to the raw data. In this study, we highlight the performance gains of FL edge caching. Subsequently, we outline the advantages and challenges of FL edge caching in vehicular Internet of things. It is shown that FL edge caching provides robust, stable, and flexible systems while ensuring a privacy-preserving solution for the data owners.**

Keywords— Edge caching; federated learning; 5G; vehicular internet of things.

I. INTRODUCTION

With the emergence of the fifth generation (5G) wireless networks, various devices in communication networks are able to connect and exchange information more efficiently than ever before [1]. Connected devices in the Internet of things (IoT) continuously generate enormous amount of data which would be requested by IoT application users. Transmitting requested IoT data from cloud servers would lead to increased network traffic and long delays. Therefore, edge caching mechanisms are used to ensure reduced latency and network traffic. Edge caching utilizes storage resources of edge nodes (ENs) by caching popular data files at the ENs. Then, user equipments (UEs) such as client vehicles in vehicular IoT can obtain the cached files from the ENs such as road side units without explicitly communicating with the cloud servers [2].

Federated learning (FL) edge caching is a decentralized machine learning (ML) edge caching technique that allows to collaboratively train a shared model for many UEs without direct access to the raw data. The technique allows distributed UEs to train local machine learning models on their local datasets and upload it to a server for a global model aggregation. FL edge caching is radically different from other more established techniques which upload raw data samples to servers.

In this study, we present an overview of the operational features of FL edge caching. Then, we highlight the performance gains of FL edge caching in 5G application scenarios. Moreover, we outline the advantages and challenges of FL edge caching in vehicular IoT.

The remainder of this paper is organized as follows. Section II presents an overview of the FL process and the unique features of the FL mechanisms. Section III outlines the advantages and challenges of FL edge caching in vehicular IoT. In section IV, the paper is concluded.

II. FEATURES OF FEDERATED LEARNING

A. Federated Learning Process

In general, FL presents an efficient training pattern which involves three main phases as demonstrated in Fig. 1. In phase 1, the global model on the server initializes model parameters and then all the UEs download the shared global model. In phase 2, each UE trains the model on its local data independently. Thus, each UE computes an individual update based on its local dataset. In phase 3, all local trained models are uploaded to the server via a secure protocol tunnel and are aggregated to learn a new global model. Then, the new global model is sent back to the UEs [3]-[5]. The learning process is iterated for many communication rounds until the global model is able to converge to the global optimal, a threshold level is achieved, or a desirable training accuracy is achieved. To ensure reduced communication overhead, FL frameworks employ selective communication such that only the important or relevant updates are transmitted in each communication round [3]. Only the model parameters or gradients are transmitted instead of the raw data.

B. Unique Features of FL

FL has the following unique properties as compared to other decentralized ML algorithms. (1) Ability to handle not independent and identically distributed (non-IID) data; the training data on a given UE is typically based on the usage of the device by a particular user, the wireless environment the UE experiences, the computation capability of the UE, and the energy consumption of the UE. Hence, any UE local dataset will not be representative of the training data of all UEs [6], [7]. In FL, the challenge of non-IID data can be met by merging the updates of the models by using the FederatedAveraging (FedAvg) algorithm [6]. (2) Unbalanced data; some users will make intensive use of a service or app than others. This can result in varying amounts of local training data. Furthermore, some UEs may have more computation tasks to be handled and some may experience more states of mobile networks. This can result in unbalancing training data among the UEs [6], [7]. Also, this challenge can be addressed by the FedAvg algorithm [6]. (3) Limited communication; UEs

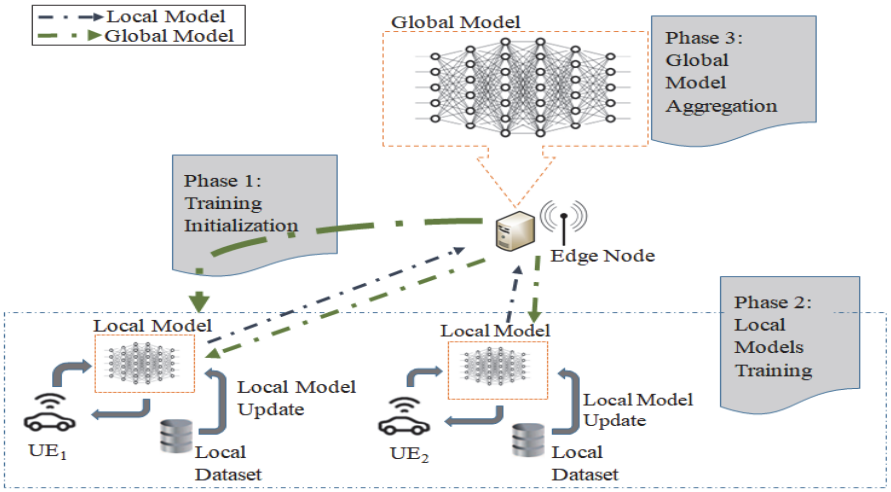

Fig. 1. An illustration of FL process.

are frequently and unpredictably offline, on slow or expensive connections, or they are allocated with poor communication resources [6], [7]. However, in FL, additional computation could decrease the consumption of communication rounds needed to train a model. Moreover, FL only asks a part of UEs, in one round, to upload their updates, which handles the situations where UEs are unpredictably offline [6]. (4) Privacy preservation; FL has the ability to collaboratively train a learning model on their individually gathered data, without revealing their privacy-sensitive data to a centralized server [5], [8]. Also, the information that needs to be uploaded to the server is the minimal update necessary. This feature is particularly useful in applications where datasets from the UEs are privacy-sensitive. Furthermore, the techniques of secure aggregation and differential privacy can be employed to ensure privacy preservation of data in the local updates [3], [6].

C. *Performance Gains of FL Edge Caching*

Several benefits of using FL in edge caching systems were demonstrated in [3], [6], [9], [10]. We summarize the benefits as follows. (1) System becomes more cognitive; in systems with a large number of UEs, the UEs can acquire various, abundant and personalized data for updating the global learning model. The data could include the quality of the wireless channel, the remaining battery life and the energy consumption, and the immediate computation capability. On the ENs, the cognitive data could include the computation load, the storage occupation, the number of wireless communication links, and the task queue states waiting for handling. As a result, the use of abundant and personalized distributed data instead of centralized training data ensures the system is more cognitive. (2) System becomes more robust; since FL can address the key issues such as the availability of the UEs, unbalanced data, and non-IID data, the performance of the systems is not easily affected by the unbalanced data or poor communication environment. Furthermore, its ability to handle non-IID data allows massive UEs in dynamic environments to train their own models without considering the overall negative effects. (3) Improved flexibility; in FL, additional computation could be used to decrease the number of communication rounds to train a model. The additional computation can be achieved by increasing the computation per UE. Therefore, UEs can decide to vary the number of mini-batches in training to adjust the communication cost. (4) Reduced network traffic and energy consumption; the decentralized training can significantly reduce the network traffic and energy consumption by sending only the features of interest rather than raw data. (5) Stability despite loss of connectivity; FL does not rely on synchronization among learners. Hence, even during a loss of connectivity between the ENs and UEs, the UEs can still build their local models. The feature is particularly important for highly dynamic and mission-critical applications such as vehicular IoT [9].

III. Federated Learning Edge Caching in Vehicular Internet of Things

A. *Suitability of FL Edge Caching in Vehicular IoT*

FL attracts a lot of interests from a large number of industries due to growing privacy concerns. Future vehicular IoT systems, such as cooperative autonomous driving and intelligent transport systems (ITS), feature a large number of devices and privacy-sensitive data where the communication, computing, and storage resources must be efficiently utilized. Therefore, FL edge caching is a promising approach to solve the existing challenges [11], [12].

The use of FL edge caching in vehicular IoT was considered in [3], [9], [11], [12], [13]. It was presented that the novel services in vehicular IoT systems present problems which can be addressed through the use of FL edge caching. The services demand unprecedented high reliability, high accuracy, and quick response. Furthermore, some services experience an extreme variance in their resource demands with respect to time, location, context, as well as individual users. Also, the vehicles are equipped with different types of sensor devices that generate and handle privacy-sensitive data, and the environments vary with time and road types. In such scenarios, FL edge caching helps to realize intelligent

vehicular IoT systems and privacy-preserving collaboration among different vehicles and road side units.

B. Challenges of FL Edge Caching in Vehicular IoT

In general, traditional FL algorithms suffer from the challenges of massive communication overhead and non-IID data [14], [15]. The accuracy of FL algorithms is significantly reduced when highly non-IID data is used as compared to when IID data is used. The communication overhead of FL mainly comes from the global model aggregation and update processes [3], [5], [8]-[12], [14], [15]. For instance, there exists FL algorithms that require each UE to communicate its full gradient update which may be in the size of gigabytes based on the learning architecture, and its millions of parameters. The size can increase to reach petabytes when the training is conducted on large-scale datasets [5]. The FedAvg is an example of FL algorithms that suffer from massive communication overhead. In [18], it was shown that the communication overhead in FL can impact other parameters such as the model accuracy and training time.

In vehicular IoT environments, the deployment of FL incurs several key challenges as highlighted in [3], [12], [13]. (1) The selection of UEs (client vehicles) for FL should address the mobility, communication bandwidth, and the specific scenarios that the UEs could represent. The mobility of vehicles makes it difficult to maintain continuous synchronized communication between the server and the vehicles. The consideration of mobility and communication bandwidth ensures a successful dissemination of the learning model and an accurate aggregation of local updates from the UEs. The consideration of vehicle scenarios guarantees that selected data for training includes a wide range of samples, avoiding the over fitting for a non-representative scenario. (2) Most vehicular IoT applications have stringent latency and reliability constraints. Therefore, it is important to devise an enhanced FL architecture which is more suitable for vehicular environments. (3) The dynamicity of vehicular environments makes the communication and computational resource allocation particularly difficult. Consequently, it becomes important to design efficient resource allocation algorithms that could satisfy the need of FL. (4) Centralized curator for aggregation is vulnerable to security threats which can lead to failure of the whole learning process. An asynchronous FL architecture was presented in [13] to mitigate the challenge.

IV. CONCLUSION

This paper presents an overview of the operational features of FL edge caching. It highlights the performance gains of FL edge caching. Furthermore, it outlines the advantages and challenges of FL edge caching in vehicular IoT. It is shown that FL frameworks exchange only model parameters learned locally at client vehicles. As a result, FL edge caching provides a better privacy-preserving solution for the data owners while ensuring robust, stable, and flexible systems.

ACKNOWLEDGMENT

This research is supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2018R1D1A1B07048338).

REFERENCES

[1] Z. Ning, et al, “Joint computing and caching in 5g-envisioned internet of vehicles: A deep reinforcement learning-based traffic control system,” *IEEE Transactions on Intelligent Transportation Systems*, Early Access Article, 2020.

[2] H. Zhu, Y. Cao, X. Wei, W. Wang, T. Jiang, and S. Jin, “Caching transient data for internet of things: A deep reinforcement learning approach,” *IEEE Internet Of Things Journal*, vol. 6, no. 2, pp. 2074–2083, April 2019.

[3] W. Y.B. Lim, et al., “Federated learning in mobile edge networks: A comprehensive survey,” *IEEE Communications Surveys & Tutorials*, Early Access Article, 2020.

[4] X. Wang, C. Wang, X. Li, V. C. M. Leung, and T. Taleb, “Federated deep reinforcement learning for internet of things with decentralized cooperative edge caching,” *IEEE Internet of Things Journal*, Early Access Article, 2020.

[5] M. Asad, A. Moustafa, and T. Ito, “FedOpt: Towards communication efficiency and privacy preservation in federated learning,” *Applied Sciences*, vol. 10, 2864, 2020.

[6] X. Wang, Y. Han, C. Wang, Q. Zhao, X. Chen, and M. Chen, “In-edge ai: Intelligentizing mobile edge computing, caching and communication by federated learning,” *IEEE Network*, vol. 33, no. 5, pp. 156–165, October 2019.

[7] H. B. McMahan, et al, “Communication-efficient learning of deep networks from decentralized data,” [Online]. Available: https://arxiv.org/abs/1602.05629, 2017.

[8] L. Wang, W. Wang, B. Li, “CMFL: Mitigating communication overhead for federated learning,” in *Proc. IEEE 39th Int. Conf. Distrib.Comput. Syst. (ICDCS)*, Jul. 2019, pp. 954–964.

[9] S. Samarakoon, M. Bennis, W. Saad, and M. Debbah, “Federated learning for ultra-reliable low-latency v2v communications,” in *Proc. 2018 IEEE Global Communications Conference (GLOBECOM)*, December 2018.

[10] S. Niknam, H. S. Dhillon, and J. H. Reed, “Federated learning for wireless communications: Motivation, opportunities and challenges,” 2020. [Online]. Available: arxiv preprint arxiv:1908.06847, 2020.

[11] J. Zhang, and K. B. Letaief, “Mobile edge intelligence and computing for the internet of vehicles,” *Proceedings of the IEEE*, vol. 108, no. 2, pp. 246–261, February 2020.

[12] Z. Du, et al, “Federated learning for vehicular internet of things: Recent advances and open issues,” *IEEE Open Journal of the Computer Society*, vol. 1, pp. 45–61, May 2020.

[13] Y. Lu, X. Huang, Y. Dai, S. Maharjan, and Y. Zhang, “Differentially private asynchronous federated learning for mobile edge computing in urban informatics,” *IEEE Transactions On Industrial Informatics*, vol. 16, no. 3, , pp. 2134–2143, March 2020.

[14] Q. Wu, K. He, and X. Chen, “Personalized federated learning for intelligent iot applications: A cloud-edge based framework,” *IEEE Open Journal of the Computer Society*, vol. 1, pp. 35–44, May 2020.

[15] Y. Sun, S. Zhou, and D. Gündüz, “Energy-aware analog aggregation for federated learning with redundant data,” in *Proc. ICC 2020*, 7-11 June 2020.

[16] L. Liu, J. Zhang, S. Song, and K. B. Letaief, “Edge-assisted hierarchical federated learning with non-iid data,” [Online]. Available: https://arxiv.org/abs/1905.06641, 2019.

[17] J. Mills, J. Hu, and G. Min, “Communication-efficient federated learning for wireless edge intelligence in iot,” *IEEE Internet Of Things Journal*, vol. 7, no. 7, pp. 5986–5994, July 2020.

[18] H. Sun, S. Li, F. R. Yu, Q. Qi, J. Wang, and J. Liao, “Towards communication-efficient federated learning in the internet of things with edge computing,” *IEEE Internet of Things Journal*, Early Access Article, May 2020.