Hierarchical Federated Learning: Architecture, Challenges, and Its Implementation in Vehicular Networks

YAN Jintao[1], CHEN Tan[1], XIE Bowen[1], SUN Yuxuan[2], ZHOU Sheng[1], NIU Zhisheng[1]

(1. Tsinghua University, Beijing 100084, China;
2. Beijing Jiaotong University, Beijing 100044, China)

DOI: 10.12142/ZTECOM.202301005

https://kns.cnki.net/kcms/detail/34.1294.TN.20230220.1541.002.html, published online February 21, 2023

Manuscript received: 2022-11-04

Abstract: Federated learning (FL) is a distributed machine learning (ML) framework where several clients cooperatively train an ML model by exchanging the model parameters without directly sharing their local data. In FL, the limited number of participants for model aggregation and communication latency are two major bottlenecks. Hierarchical federated learning (HFL), with a cloud-edge-client hierarchy, can leverage the large coverage of cloud servers and the low transmission latency of edge servers. There are growing research interests in implementing FL in vehicular networks due to the requirements of timely ML training for intelligent vehicles. However, the limited number of participants in vehicular networks and vehicle mobility degrade the performance of FL training. In this context, HFL, which stands out for lower latency, wider coverage and more participants, is promising in vehicular networks. In this paper, we begin with the background and motivation of HFL and the feasibility of implementing HFL in vehicular networks. Then, the architecture of HFL is illustrated. Next, we clarify new issues in HFL and review several existing solutions. Furthermore, we introduce some typical use cases in vehicular networks as well as our initial efforts on implementing HFL in vehicular networks. Finally, we conclude with future research directions.

Keywords: hierarchical federated learning; vehicular network; mobility; convergence analysis

Citation (IEEE Format): J. T. Yan, T. Chen, B. W. Xie, and et al., "Hierarchical federated learning: architecture, challenges, and its implementation in vehicular networks," *ZTE Communications*, vol. 21, no. 1, pp. 38–45, Mar. 2022. doi: 10.12142/ZTECOM.202301005.

1 Introduction

Recently, the evolution of intelligent technologies gives rise to a wide range of emerging applications including the Internet of Things (IoT), autonomous driving, and so on. While opening up new ways of life for users, these applications also produce numerous data scattered on mobile devices. Transmitting these data to a centralized server for traditional machine learning (ML) is no longer capable due to limited communication resources, tight latency requirements and stringent privacy concerns. As a result, federated learning (FL) is proposed as a distributed learning solution, where multiple mobile devices and a parameter server cooperatively train an ML model by only exchanging the model parameters without directly sharing their local data.

In recent years, many works have been done to deal with the different challenges of FL[1]. Among them, communication efficiency is one of the most important issues[2]. Many FL frameworks consider the cloud server as the parameter server, but the communication between clients and the cloud server is inefficient and unpredictable. Federated edge learning (FEEL)[3], where the clients share the ML model parameters with edge servers, has been proposed to reduce communication latency. However, the edge servers in FEEL have limited coverage and the number of clients for FL training cannot meet the requirements, resulting in the degradation of training performance. Therefore, it is necessary to characterize the tradeoff between communication latency and training performance.

To deal with this issue, the concept of hierarchical FL (HFL) has been proposed[4–5], which leverages the large coverage of the cloud server and the high communication efficiency of the edge server. This architecture consists of one cloud server, multiple edge servers, and a multitude of clients. In HFL, the clients update their local parameters and send them to the edge servers for edge aggregations as conventional FL does. The difference is that after several rounds of edge aggregations, multiple edge servers send their parameters to a cloud

The work of YAN Jintao, CHEN Tan, XIE Bowen, ZHOU Sheng and NIU Zhisheng are sponsored in part by the National Key R&D Program of China under Grant No. 2020YFB1806605, the National Natural Science Foundation of China under Grant Nos. 62022049, 62111530197, and 61871254, and OPPO. The work of SUN Yuxuan is supported by the Fundamental Research Funds for the Central Universities under Grant No. 2022JBXT001.

server for cloud aggregation, which allows more clients to be involved in the framework. Experimental results and theoretical analysis have shown that this client-edge-cloud FL architecture has higher convergence speed and less training time compared with the conventional framework[5].

During the last several years, FL has witnessed its potential in vehicular networks. The advantage of implementing FL in vehicular networks is twofold. First, FL can satisfy the latency and privacy requirements that the applications in vehicular networks, such as trajectory planning and traffic flow optimization, call for. Second, intelligent vehicles have computation and communication capabilities and can sample abundant data for training[6]. There have been many papers on the implementation of FL in vehicular networks. In Ref. [7], an FL-based approach is proposed to allocate the power and resource for ultra-reliable low-latency communications in vehicular networks. In Ref. [8], FL is used to update an edge caching scheme for vehicular networks, which considers the cached content and vehicular mobility. Considering the computation and communication resources and local dataset of vehicles, the authors of Ref. [9] propose a joint vehicle selection and resource allocation scheme for FL training. In Ref. [10], the vehicle speed and position are taken into consideration and an optimization problem is formulated for resource allocation for FL.

However, implementing FL in vehicular networks may be more challenging than that in conventional wireless networks[11]. First, the ML application in vehicular networks has more stringent requirements for latency. This is because vehicles may leave the coverage of the central server due to mobility before successfully uploading their updated local model to the server. Second, since the physical distances between vehicles are much larger than those of humans or mobile devices, the number of clients participating in model aggregation in vehicular networks is much lower than that in conventional wireless networks, which degrades the convergence performance of FL. In this context, HFL stands out for its properties of lower latency, wider coverage, and more participants, inspiring us to search for the possibility of implementing HFL in vehicular networks.

There are some existing surveys and tutorials on FL, as shown in Table 1. In Ref. [1], a comprehensive survey of FL in wireless networks is provided, and research directions including compression and sparsification, convergence analysis, wireless resource management, and FL training method design are presented. The authors of Ref. [2] focus on minimizing the communication and computation latency and introduce the concept of timely edge learning. The key challenges and solutions to the timely issues are discussed. In Ref. [11], the concept of FL is combined with mobile edge computing (MEC) and a comprehensive survey of FL and MEC is provided. In Ref. [12], the implementation of FL in vehicular networks is studied and the major challenges are analyzed from a learning and communication perspective. In this work, we provide a comprehensive review of HFL and explore the feasibility of implementing HFL in vehicular networks.

▼Table 1. Existing surveys on FL

Highlight	Reference
A comprehensive survey of FL in wireless networks	Ref. [1]
A tutorial on timely edge learning, aiming to minimize the communication and computation latency in FL training	Ref. [2]
A comprehensive survey of FL and MEC	Ref. [11]
A tutorial on the implementation of FL in vehicular networks and the major challenges of learning and communications	Ref. [12]

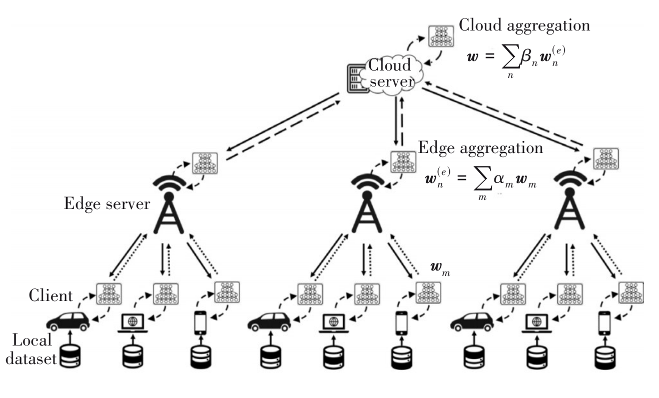
FL: federated learning MEC: mobile edge computing

The rest of this paper is organized as follows. The HFL architecture is introduced in Section 2. In Section 3, we clarify the new issues and challenges in HFL compared with FL and provide a review of existing works dealing with these issues. In Section 4, we introduce the typical use cases of HFL in vehicular networks. Section 4 concludes this paper and gives some future research directions in this field.

2 HFL Architecture

In a typical HFL system, a cloud, some edges and several clients collaboratively train an ML model. The cloud covers all the edges and each edge covers some of the clients. All of the participants initialize a model of the same parameters and perform cloud epochs. Each cloud epoch is composed of edge learning stages and a cloud aggregation stage. During the edge learning stage, each edge, together with clients under its coverage, trains the learning model in the way of FL for some iterations. During the cloud aggregation stage, edges transmit their model parameters or gradients to the cloud. The cloud aggregates the parameters or gradients to update the global model and broadcasts the global model to edges. The cloud epoch is repeated until the global model converges. The training procedure is illustrated in Fig. 1.

Different from clients in FL, who are always connected to the same parameter server, those clients in HFL can be associated with different edges during training. First, in cellular net-



▲Figure 1. Architecture of a hierarchical federated learning system

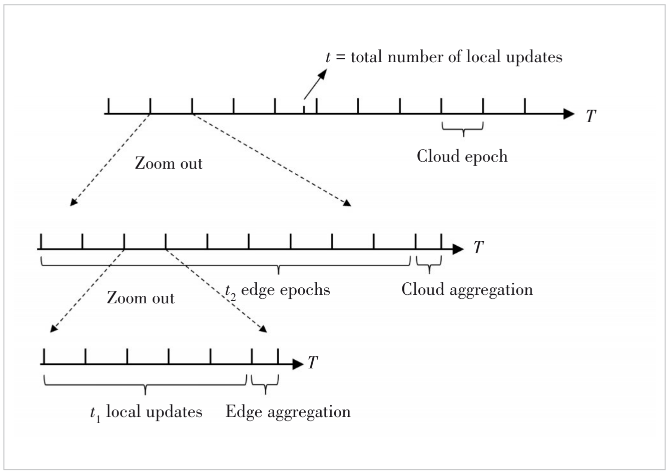
works, the coverage areas of cells are generally overlapping, so clients in the overlapped area of some edges can be associated with either of them. Second, in a scenario of wireless communications, especially in vehicular networks, clients may be moving, which means they can step from the coverage of one edge into that of another during training, while generally staying in the range of the cloud. Therefore, in HFL, edges may need to reconstruct connections with clients at the beginning of each iteration.

To formulate the training procedure, we assume there are M clients and N edges and denote $\boldsymbol{w}_m(t)$ as a client's m-th local model parameters at the t-th local update. Assume the clients perform local updates t_1 before edge aggregation and the edges perform FL iterations t_2 before cloud aggregation. For client m, given the loss function $F_m(\cdot)$, learning rate η and the set of clients that are associated with the same edge at the t-th local update $\varepsilon_m^{(t)}$, the local model evolves as follows:

$$\tilde{\boldsymbol{w}}_m(t) = \boldsymbol{w}_m(t-1) - \eta\nabla F_m(\boldsymbol{w}_m(t-1)),$$

$$\boldsymbol{w}_m(t) = \begin{cases} \tilde{\boldsymbol{w}}_m(t), t_1 \nmid t \\ \sum_{i \in \varepsilon_m^{(t)}} \alpha_i \tilde{\boldsymbol{w}}_i(t), t_1 | t, t_1 t_2 \nmid t \\ \sum_{j=1}^{N} \beta_j \sum_{i \in \varepsilon_m^{(t)}} \alpha_i \tilde{\boldsymbol{w}}_i(t), t_1 t_2 | t \end{cases},$$

where α_i and β_j are edge and cloud aggregation weights separately, with $\sum_{i \in \varepsilon_m^{(t)}} \alpha_i = 1$ and $\sum_j \beta_j = 1$. Here $a|b$ means b is divisible by a. On the opposite, $a \nmid b$ means b is not divisible by a. The timescale of HFL training is further shown in Fig. 2, which demonstrates the relationships of t, t_1 and t_2 more clearly.



▲Figure 2. Timescale of hierarchical federated learning (HFL) training

3 Overview of New Research Issues in HFL

Compared with FL, HFL brings many new research issues, both theoretically and practically. From the theoretical perspective, the convergence analysis for HFL is more complex because of the multi-layer architecture. From the practical perspective, the resource management strategies for HFL should not only focus on allocating the wireless resources under one server, but also arrange resources among different edge servers. Also, the popularity of HFL gives rise to many new considerations, such as HFL with device-to-device (D2D) communications and mobility-aware HFL. We provide a survey on HFL based on these three categories: convergence analysis, resource management, and new considerations of HFL. Note that these three categories might overlap with each other. For instance, the convergence analysis results might be used to design the resource allocation strategy in some works.

3.1 Convergence Analysis

In FL, convergence analysis illustrates how different factors influence the FL training performance, and thus can be used as a guideline for FL system design. In HFL, the convergence analysis is more complex. For FL, the clients only perform local updates before global aggregation. However, edge aggregation is conducted before global aggregation, which results in a loose bound of convergence analysis. Many works on convergence analysis for HFL have been done. In Ref. [5], an HFL framework is proposed and the convergence analysis of this framework is provided. By investigating how the distributed weights deviate from the centralized sequence, the authors give an upper bound for the deviation. The results show how the edge and the cloud intervals influence the convergence performance for both convex and non-convex loss functions. Following this work, the authors of Ref. [13] provide a tighter convergence bound. In this work, model quantization is adopted to improve communication efficiency, and the edge and cloud aggregation intervals are optimized based on the theoretical results to improve the training performance. The authors of Ref. [14] assume a graph topology where each edge is considered as a node in the graph, and occasionally averages its model parameters with adjacent nodes in a decentralized manner. Furthermore, a probabilistic approach is adopted for analyzing local updates. Convergence analysis of this scenario is then provided, showing the influence of local iterations, edge epochs, cloud epochs, network topology and node heterogeneity on the convergence performance. Ref. [15] is the first work that takes both data heterogeneity and stochastic gradient descent into consideration for convergence analysis. By denoting client-edge and edge-cloud data divergence, data heterogeneity is connected to the convergence bound and a worst-case upper bound for convergence is provided. The convergence bound shows that local aggregates accelerate the convergence speed of the global model by a "sandwich" behavior. The results are also extended to the cases in which the grouping is random or there are more than three layers.

However, most of the above papers consider a static topology. In vehicular networks, the mobility of clients may degrade model convergence, which should be taken into consideration. The authors of Ref. [16] propose a mobility-aware HFL framework. First, the HFL framework with mobile clients is modeled by a Markov chain. Then, convergence analysis is provided, showing how user mobility influences training performance. Based on the theoretical analysis, the local update mode and access scheme are modified to reduce the impact of client mobility. Experimental results illustrate that the proposed scheme can outperform the baselines, especially when the data heterogeneity or user mobility is high or the number of users is small.

3.2 Resource Management

Resource management is an important issue in FL. It means how the communication bandwidth, power and computing resources are allocated to clients under the coverage of one server. In HFL, there is more than one edge server and new issues arise.

One new issue in HFL is edge association, which is defined to find which clients should be allocated to which edge server. In Ref. [17], a joint resource allocation and edge association problem is formulated under HFL. The authors first propose the architecture of HFL and an optimization problem that aims to minimize both latency and energy consumption. Then, this problem is decomposed into two subproblems: a resource allocation problem and an edge association one. The resource allocation problem is proved to be convex and the optimization value can be reached. The edge association problem is solved via an iterative global cost reduction adjustment method. Simulation results show that the proposed scheme can outperform the baselines in terms of FL training performance with low latency and energy consumption. The authors of Ref. [18] focus on the interactions and limited rationalities of the clients. A dynamic resource allocation and edge association problem is proposed based on the game theory in self-organizing HFL frameworks. The edge association problem is solved via a lower-level evolutionary game and the resource allocation problem is solved via an upper-level Stackelberg differential game. Experiments show that the proposed scheme can well suit the dynamics of the HFL system. In Ref. [19], the effect of data heterogeneity is taken into consideration. The model error and the latency for HFL are first analyzed, and the optimization problem of user association and resource allocation is then proposed under both independent identically distributed (i.i.d.) and non-i.i.d. settings. For the non-i.i.d. settings, the distance of data distribution is considered and a primal-dual algorithm is proposed to solve the problem. Simulation results show that under both i.i.d. and non-i.i.d. settings, the proposed scheme can outperform the baselines in terms of latency and testing accuracy.

Other issues in HFL include aggregation interval and incentive mechanism design. In Ref. [20], a joint resource allocation and aggregation interval control problem is proposed, aiming to minimize the training loss and the latency. Convergence analysis is provided to show the dependency of the convergence performance on the number of participants, the aggregation interval and training latency. Then, the original problem is decomposed into two subproblems. The resource allocation problem is proved to be convex and the optimal value can be reached. For the aggregation interval control problem, a rounding and relaxation approach is adopted. Experimental results show that the proposed scheme can reach lower latency and higher training performance compared with the baselines. In Ref. [21], a two-level joint incentive design and resource allocation problem is proposed. At the lower level, the cluster selection problem is formulated as an evolutionary game. At the upper level, the action of the cluster head is solved via a deep learning-based approach. Experiments show the robustness and uniqueness of the proposed scheme.

3.3 New Considerations of HFL

The popularity of HFL gives rise to many novel architectures, such as HFL with device-to-device (D2D) communications. In Ref. [22], a multi-layer hybrid FL framework is proposed. The authors first introduce the architecture of this new FL architecture, where there are more than three layers. In each layer, clients aggregate the model parameters via D2D

▼Table 2. Summary of recent papers on HFL

Category	Highlight	Reference
Convergence analysis	Effect of edge and cloud aggregation intervals and local update step size with both convex and non-convex loss functions	Ref. [5]
	Extending Ref. [5] into HFL with quantization and carrying out the convergence analysis	Ref. [13]
	Effect of local iterations, edge epochs, global epochs, network topology and node heterogeneity on the convergence performance for a graph-based edge topology	Ref. [14]
	Extending Ref. [14] into HFL with data heterogeneity, random grouping and multi-layer architecture	Ref. [15]
	Mobility-aware HFL	Ref. [16]
Resource management	Joint resource allocation and edge association	Refs. [17 – 19]
	Joint resource allocation and interval control	Ref. [20]
	Joint resource allocation and incentive mechanism design	Ref. [21]
Other practical considerations	Multi-stage HFL with device-to-device communications	Ref. [22]

HFL: hierarchical federated learning

communications and then transmit the parameters to the upper layers. Convergence analysis is provided to derive an upper bound of this framework and a distributed control algorithm is proposed to improve the convergence performance. Experiment results show that the proposed framework can utilize the network resources more efficiently without loss of convergence speed and testing accuracy.

4 HFL in Vehicular Networks

There are many application scenarios that can benefit from the deployment of HFL in vehicular networks, such as autonomous driving, intelligent transportation systems and smart wireless communications. Recent studies on these scenarios have adopted FL as the training framework of AI models to obtain advantages in higher convergence speed, lower energy consumption and better privacy protection[23]. However, research on applying HFL to vehicular networks is still in its infancy, leaving a large room for further study.

In this section, we first introduce several typical use cases of ML in vehicular networks, showing the great potential of HFL. Then we analyze the challenges and opportunities of HFL caused by mobility in vehicular networks. Finally, we show our own work on the implementation of HFL in vehicular networks, taking into account the mobility aspect.

4.1 Typical Use Cases

1) Autonomous driving: Autonomous driving is one of the key technologies in future vehicular networks. Trajectory prediction and path planning are two necessary capabilities of autonomous driving vehicles. To avoid collision with pedestrians, vehicles and other traffic agents, autonomous driving vehicles must reliably predict the future trajectories of surrounding agents and safely and efficiently plan their own future driving paths[24]. Their decisions are based on the sensing data from onboard cameras, Lidars, GPS, and map information. To meet the stringent latency and precision requirements, ML algorithms have been applied for these two tasks[25 – 26], which perform better than traditional approaches. However, the traffic environments of vehicular networks vary all the time as they keep driving, which requires vehicles to continually update their ML models with the latest data generated by sensors. HFL is more promising to provide well-trained and up-to-date ML models over centralized ML or conventional FL, since HFL can utilize much more training data generated from a large number of vehicles driving in various areas, which can improve the adaptability of ML models to dynamic environments.

2) Intelligent transportation systems (ITS): ITS are novel traffic systems that utilize advanced information technologies to reduce traffic congestion, accident rate, energy consumption and carbon emissions, and thus enhance efficiency, safety, reliability and eco-friendliness[27]. Many typical applications of ITS are critical to future vehicular networks, such as collaborative perception and vehicle platooning. Collaborative perception, where data from multiple traffic agents are collected and fused to conduct object detection, can achieve higher accuracy and precision than single-vehicle perception[28]. Vehicle platooning, where a coordinated group of autonomous vehicles travels collectively, can achieve faster and safer autonomous driving with shorter spacing than single-vehicle traveling[29]. Existing research[28, 30] on these use cases also considers applying machine learning methods to achieve better performance. Note that the ML models for ITS tasks usually require vehicles to share data, and the data, such as photographs and videos, can be private and sensitive. However, the centralized ML needs to collect the raw data from all vehicles to train an ML model, which leads to heavy communication burdens, as well as privacy problems. To reduce the unnecessary raw data transmission and the resulting privacy leakage, HFL is a promising paradigm of model training in ITS, since it only collects the lightweight gradient data, rather than the heavyweight and private raw data.

3) Smart wireless communications: In smart wireless communications, ML algorithms are utilized in many wireless communication tasks, such as multiple - input, multiple - output (MIMO) beam selection[31], channel modeling and estimation[32], and joint source-channel coding[33]. Compared to traditional wireless communications, ML algorithms designed and exploited for smart wireless communications can decrease communication overhead, improve the signal-to-noise ratio (SNR), and save transmission power, with much lower latency and fewer computing resources. Similar to the use cases of autonomous driving, it is a challenge for ML models to adapt to the dynamic characteristics of channel states in vehicular networks. Therefore, HFL is also a promising training approach for smart wireless communications.

Although the use cases aforementioned have taken ML into account, there are few papers applying HFL to train the ML models for these scenarios in vehicular networks. Actually, HFL can exploit the data and computing resources of more vehicles, and thus train ML models more efficiently than centralized ML. Compared to FL, vehicles from larger areas can bring richer data features to the training of HFL, which improves the robustness of ML models. Therefore, it is promising to further study the application of HFL in vehicular networks.

4.2 Challenges and Opportunities with Vehicle Mobility

Despite the promising potential of applying HFL to vehicular networks, some properties of vehicular networks may stand as great barriers, in particular the mobility. Unlike other FL scenarios where clients stay in the same place or move at a low speed, intelligent vehicles usually travel fast on road, especially when they drive on the highway. This brings more dynamics and uncertainties to the topology of vehicles, leading to a change of association between vehicles and edges. First, vehicles may leave the coverage of an edge when uploading its model parameters while transmitting model parameters, or even before finishing one round of local updates, leading to a waste of communication and computation resources as well as leakage of training data. Second, the varying channel conditions of vehicular communication links and the Doppler effect caused by vehicle mobility may result in the failure of model transmission or transmission errors in the received parameters, which also influences the FL training performance.

However, there are also chances brought by mobility. On the one hand, the mobility of vehicles creates more opportunities to meet[6] other vehicles, inspiring the leverage of vehicle-to-vehicle (V2V) communications through side links to compensate for the loss of changing edge and also accelerate the speed of edge aggregation. On the other hand, since the hierarchical structure of HFL brings a wide coverage, even though vehicles step out of the coverage of an edge, there's a great chance that they still stay in the range of the cloud, so their data can still be used by training. What's more, due to the heterogeneity of clients and the dynamic nature of the road environment, data distribution generally varies from one edge to another. Mobility of vehicles promotes data fusion of edges and thus reduces data heterogeneity, which helps the global training model to converge faster. In the following section, we will give two case studies as examples of leveraging these opportunities.

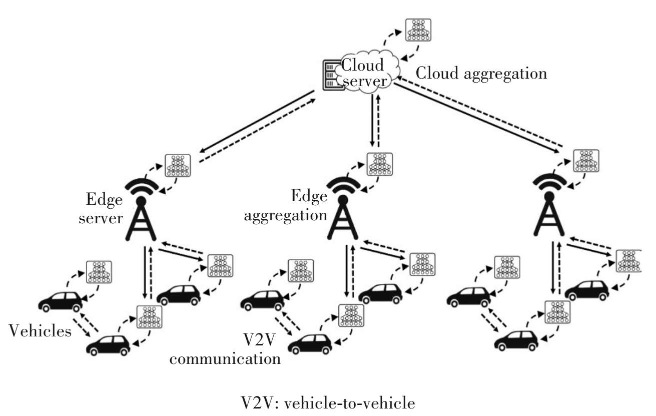
4.3 Case Study 1: V2V-Assisted Hierarchical Federated Learning

In this case study, we propose a V2V-assisted hierarchical federated learning (VAHFL) framework, where the V2V communication is utilized to speed up the aggregation process. In this framework (Fig. 3), the uploading of model parameters includes both vehicle-to-infrastructure (V2I) and V2V communication. Some vehicles act as relay nodes that help other vehicles with parameter transmission. Vehicles leaving the coverage of the central server can transmit their model parameters to the nearby relay nodes via the V2V link before it leaves, while vehicles near the server directly transmit its parameter to the server via the V2I link. We formulate a communication latency minimization problem by optimizing the uploading strategy, and a graph neuron network-reinforcement learning (GNN-RL) based algorithm is designed to solve this problem.

An experimental platform is built based on Simulation of Urban Mobility (SUMO) to evaluate the proposed framework, where there is one cloud server, four edge servers and 200 vehicles. The vehicles move over time according to the Manhattan mobility model. The vehicles cooperatively train a convolutional neural network (CNN) model for an image classification task using the CIFAR-10[34] dataset. The V2I bandwidth is set to 30 MHz, and the V2V bandwidth is set to 10 MHz. For the benchmark, we consider that the vehicles directly transmit their model parameters to the server. Fig. 4 illustrates that the proposed framework can reduce transmission latency by 41.54% and increase the percentage of successful transmissions by 10.97%.

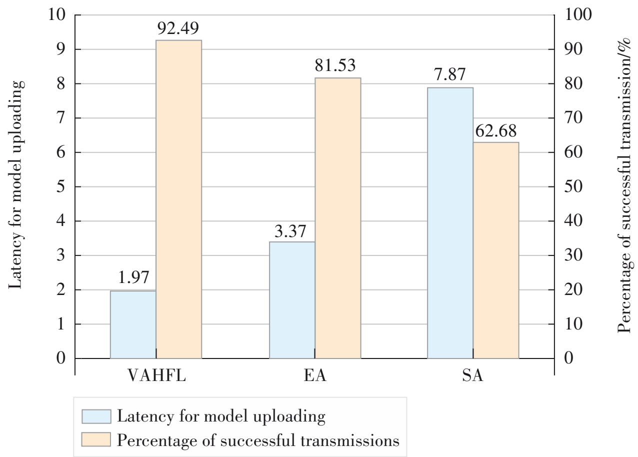
4.4 Case Study 2: Edge-Heterogeneous Hierarchical Federated Learning

In this case study, we investigate the influence of mobility when training data of edge servers are heterogeneous. Before training, vehicles sample data to form local datasets. Data distribution is dependent on the location of vehicles, which means vehicles under the coverage of the same edge server sample from the same distribution, while vehicles from the coverage of different edge servers sample differently. Therefore, at the start of training, the data distribution of edges is heterogeneous. During training, vehicles constantly travel across edges, driving the data from different edge servers to mix up. We analyze the convergence speed of this edge-heterogeneous HFL system and prove that mobility accelerates convergence by promoting data fusion.



V2V: vehicle-to-vehicle

▲ **Figure 3. Schematic of V2V-assisted hierarchical federated learning framework**



VAHFL: V2V-assisted hierarchical federated learning framework
EA: evolutionary algorithm
SA: simulated annealing

▲ **Figure 4. Latency and the percentage of successful transmission of the proposed scheme and baseline**

Experiments are also conducted based on SUMO. We assume one cloud server, four edge servers and 32 vehicles cooperatively train a four-layer CNN on the CIFAR-10 dataset, and we only choose data of eight classes from 10 classes for training and inference. Initially, each edge has data of two classes, which is uniformly distributed in vehicles under the coverage of the edge. During training, vehicles travel by the Manhattan mobility model, with their local datasets unchanged, which leads to changes in edge data distribution. The network is trained on three settings of vehicle mobility: no mobility, low mobility and high mobility. As Fig. 5 shows, mobility increases the convergence speed and final test accuracy of HFL. What' s more, when vehicles are moving, a higher vehicle speed results in a faster convergence speed. As is shown by the dashed line and stars in the figure, if we set the target test accuracy as 0.75, the low mobility and high mobility scenario reduces the training epochs by 40.6% and 51.9% separately.

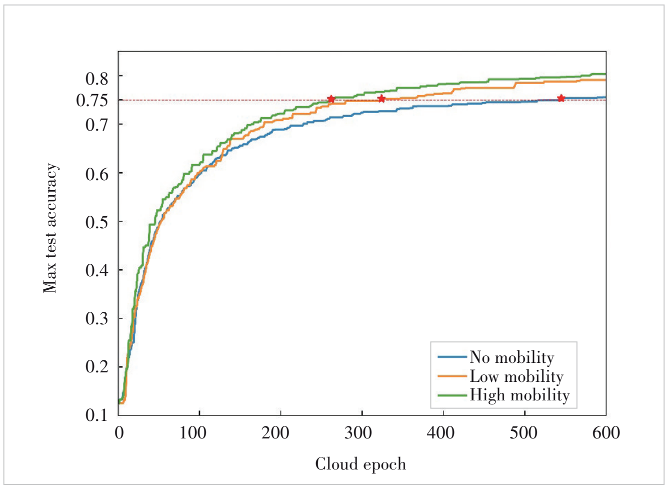
5 Conclusions

This paper presents an overview of HFL and its application in vehicular networks. First, we introduce the background and motivation of HFL and the possibility of implementing it in vehicular networks. Then, the architecture of HFL is presented. Afterward, we discuss new issues and challenges of HFL compared with FL and review existing solutions. Furthermore, some typical use cases in vehicular networks are introduced and our existing works of implementing HFL in vehicular networks are presented. Apart from the works mentioned above, there are still some challenges and research directions for HFL and its implementation in vehicular networks:

1) Heterogenous vehicular networks: For HFL in vehicular networks, the participants may be more than just vehicles. Mobile devices and other transportation infrastructures can also participate in model aggregation. In such a case, the network is heterogeneous, i.e., the computing capability, the communication capacity and the mobility patterns of clients in this network are quite different. This brings challenges to FL system design and resource management strategy.

2) Variation of channel conditions: Due to the high mobility of vehicles, the channel conditions of vehicular communication links may vary rapidly. This may result in the failure of model transmission or transmission errors in the received parameters. Therefore, the communication system should be carefully designed to prevent such cases.

3) Exploration of benefits of mobility: Usually, mobility is considered a bottleneck for FL implementation and training. However, mobility may also be explored to enhance FL training performance. In our initial efforts, the convergence speed of an edge-heterogeneous HFL is shown to be enhanced by the data fusion brought by vehicle mobility. Apart from that, other benefits of utilizing vehicle mobility are also worth being explored.



▲Figure 5. Maximum achievable test accuracy of cloud model with different mobility

References

[1] CHEN M Z, GÜNDÜZ D, HUANG K B, et al. Distributed learning in wireless networks: recent progress and future challenges [J]. IEEE journal on selected areas in communications, 2021, 39(12): 3579 – 3605. DOI: 10.1109/JSAC.2021.3118346

[2] SUN Y X, SHI W Q, HUANG X F, et al. Edge learning with timeliness constraints: challenges and solutions [J]. IEEE communications magazine, 2020, 58 (12): 27 – 33. DOI: 10.1109/MCOM.001.2000382

[3] SHI Y M, YANG K, JIANG T, et al. Communication-efficient edge AI: algorithms and systems [J]. IEEE communications surveys & tutorials, 2020, 22(4): 2167 – 2191. DOI: 10.1109/comst.2020.3007787

[4] ABAD M S H, OZFATURA E, GUNDUZ D, et al. Hierarchical federated learning across heterogeneous cellular networks [C]//IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020: 8866 – 8870. DOI: 10.1109/ICASSP40776.2020.9054634

[5] LIU L M, ZHANG J, SONG S H, et al. Client-edge-cloud hierarchical federated learning [C]//IEEE International Conference on Communications (ICC). IEEE, 2020: 1 – 6. DOI: 10.1109/ICC40277.2020.9148862

[6] SUN Y X, XIE B W, ZHOU S, et al. MEET: mobility-enhanced edge intelligence for smart and green 6G networks [J]. IEEE communications magazine, 2023, 61(1): 64 – 70. DOI: 10.1109/MCOM.001.2200252

[7] SAMARAKOON S, BENNIS M, SAAD W, et al. Distributed federated learning for ultra-reliable low-latency vehicular communications [J]. IEEE transactions on communications, 2020, 68(2): 1146 – 1159. DOI: 10.1109/TCOMM.2019.2956472

[8] YU Z X, HU J, MIN G Y, et al. Mobility-aware proactive edge caching for connected vehicles using federated learning [J]. IEEE transactions on intelligent transportation systems, 2021, 22(8): 5341 – 5351. DOI: 10.1109/TITS.2020.3017474

[9] XIAO H Z, ZHAO J, PEI Q Q, et al. Vehicle selection and resource optimization for federated learning in vehicular edge computing [J]. IEEE transactions on intelligent transportation systems, 2022, 23(8): 11073 – 11087. DOI: 10.1109/TITS.2021.3099597

[10] WANG S Y, LIU F F, XIA H L. Content-based vehicle selection and resource allocation for federated learning in IoV [C]//IEEE Wireless Communications and Networking Conference Workshops (WCNCW). IEEE, 2021: 1 – 7. DOI: 10.1109/WCNCW49093.2021.9419986

[11] LIM W Y B, LUONG N C, HOANG D T, et al. Federated learning in mobile edge networks: a comprehensive survey [J]. IEEE communications surveys & tutorials, 2020, 22(3): 2031 – 2063. DOI: 10.1109/COMST.2020.2986024

[12] ELBIR A M, SONER B, ÇÖLERI S, et al. Federated learning in vehicular networks [C]//IEEE International Mediterranean Conference on Communications

and Networking (MeditCom). IEEE, 2022: 72 – 77. DOI: 10.1109/MeditCom55741.2022.9928621

[13] LIU L M, ZHANG J, SONG S H, et al. Hierarchical federated learning with quantization: convergence analysis and system design [J]. IEEE transactions on wireless communications, 2023, 22(1): 2 – 18. DOI: 10.1109/TWC.2022.3190512

[14] CASTIGLIA T, DAS A, PATTERSON S. Multi-level local SGD: distributed SGD for heterogeneous hierarchical networks [C/OL]. International Conference on Learning Representations, 2021 [2021-05-03]. https://openreview.net/pdf?id=C70cp4Cn32

[15] WANG J Y, WANG S Q, CHEN R R, et al. Demystifying why local aggregation helps: convergence analysis of hierarchical SGD [J]. Proceedings of the AAAI conference on artificial intelligence, 2022, 36(8): 8548 – 8556. DOI: 10.1609/aaai.v36i8.20832

[16] FENG C, YANG H H, HU D, et al. Mobility-aware cluster federated learning in hierarchical wireless networks [J]. IEEE transactions on wireless communications, 2022, 21(10): 8441 – 8458. DOI: 10.1109/TWC.2022.3166386

[17] LUO S Q, CHEN X, WU Q, et al. HFEL: joint edge association and resource allocation for cost-efficient hierarchical federated edge learning [J]. IEEE transactions on wireless communications, 2020, 19(10): 6535 – 6548. DOI: 10.1109/TWC.2020.3003744

[18] LIM W Y B, NG J S, XIONG Z H, et al. Dynamic edge association and resource allocation in self-organizing hierarchical federated learning networks [J]. IEEE journal on selected areas in communications, 2021, 39(12): 3640 – 3653. DOI: 10.1109/JSAC.2021.3118401

[19] LIU S L, YU G D, CHEN X F, et al. Joint user association and resource allocation for wireless hierarchical federated learning with non-IID data [C]//IEEE International Conference on Communications. IEEE, 2022: 74 – 79. DOI: 10.1109/ICC45855.2022.9839164

[20] XU B, XIA W C, WEN W L, et al. Adaptive hierarchical federated learning over wireless networks [J]. IEEE transactions on vehicular technology, 2022, 71(2): 2070 – 2083. DOI: 10.1109/tvt.2021.3135541

[21] LIM W Y B, NG J S, XIONG Z H, et al. Decentralized edge intelligence: a dynamic resource allocation framework for hierarchical federated learning [J]. IEEE transactions on parallel and distributed systems, 2022, 33(3): 536 – 550. DOI: 10.1109/TPDS.2021.3096076

[22] HOSSEINALIPOUR S, AZAM S S, BRINTON C G, et al. Multi-stage hybrid federated learning over large-scale D2D-enabled fog networks [J]. IEEE/ACM transactions on networking, 2022, 30(4): 1569 – 1584. DOI: 10.1109/TNET.2022.3143495

[23] DU Z Y, WU C, YOSHINAGA T, et al. Federated learning for vehicular Internet of Things: recent advances and open issues [J]. IEEE open journal of the computer society, 2020, 1: 45 – 61. DOI: 10.1109/OJCS.2020.2992630

[24] MA Y X, ZHU X G, ZHANG S B, et al. TrafficPredict: trajectory prediction for heterogeneous traffic-agents [J]. Proceedings of the AAAI conference on artificial intelligence. AAAI, 2019, 33(1): 6120 – 6127. DOI: 10.1609/aaai.v33i01.33016120

[25] ALTCHÉ F, DE LA FORTELLE A. An LSTM network for highway trajectory prediction [C]//IEEE 20th International Conference on Intelligent Transportation Systems (ITSC). IEEE, 2018: 353 – 359. DOI: 10.1109/ITSC.2017.8317913

[26] SHALEV-SHWARTZ S, SHAMMAH S, SHASHUA A. Safe, multi-agent, reinforcement learning for autonomous driving [EB/OL]. (2016-10-11)[2022-05-01]. https://arxiv.org/abs/1610.03295

[27] QURESHI K N, ABDULLAH A H. A survey on intelligent transportation systems [J]. Middle-east journal of scientific research, 2013, 15(5): 629 – 642. DOI: 10.5829/idosi.mejsr.2013.15.5.11215

[28] MAO R, GUO J, JIA Y, et al. DOLPHINS: dataset for collaborative perception enabled harmonious and interconnected self-driving [EB/OL]. (2022-07-15) [2022-08-01]. https://arxiv.org/abs/2207.07609

[29] AXELSSON J. Safety in vehicle platooning: a systematic literature review [J]. IEEE transactions on intelligent transportation systems, 2017, 18(5): 1033 – 1045. DOI: 10.1109/TITS.2016.2598873

[30] PRATHIBA S B, RAJA G, DEV K, et al. A hybrid deep reinforcement learning for autonomous vehicles smart-platooning [J]. IEEE transactions on vehicular technology, 2021, 70(12): 13340 – 13350. DOI: 10.1109/TVT.2021.3122257

[31] KLAUTAU A, BATISTA P, GONZÁLEZ-PRELCIC N, et al. 5G MIMO data for machine learning: application to beam-selection using deep learning [C]//Information Theory and Applications Workshop (ITA). IEEE, 2018: 1 – 9. DOI: 10.1109/ITA.2018.8503086

[32] ALDOSSARI S M, CHEN K C. Machine learning for wireless communication channel modeling: an overview [J]. Wireless personal communications, 2019, 106(1): 41 – 70. DOI: 10.1007/s11277-019-06275-4

[33] KURKA D B, GÜNDÜZ D. Bandwidth-agile image transmission with deep joint source-channel coding [J]. IEEE transactions on wireless communications, 2021, 20(12): 8081 – 8095. DOI: 10.1109/TWC.2021.3090048

[34] KRIZHEVSKY A. Learning multiple layers of features from tiny images [D]. Toronto: University of Toronto, 2009

Biographies

YAN Jintao is a PhD student at Tsinghua University, China. His research interests include federated learning and vehicular edge computing and vehicular networks.

CHEN Tan is a PhD student at Tsinghua University, China. His research interests include federated learning and vehicular networks.

XIE Bowen is a PhD student at Tsinghua University, China. His research interests include federated learning and vehicular networks.

SUN Yuxuan is an associate professor with the School of Electronic and Information Engineering, Beijing Jiaotong University, China and was previously a postdoctoral researcher with Tsinghua University, China. Her research interests include edge computing and edge learning.

ZHOU Sheng (sheng.zhou@tsinghua.edu.cn) is an associate professor with the Department of Electronic Engineering, Tsinghua University, China. His research interests include vehicular networks, mobile edge computing, and green wireless communications.

NIU Zhisheng is a professor with the Department of Electronic Engineering, Tsinghua University, China. His major research interests include queueing theory, traffic engineering, radio resource management of wireless networks, and green communication and networks.