Federated Learning for 6G: A Survey From Perspective of Integrated Sensing, Communication and Computation

ZHAO Moke, HUANG Yansong, LI Xuan

(Beijing University of Posts and Telecommunications, Beijing 100876, China)

DOI: 10.12142/ZTECOM.202302005

https://kns.cnki.net/kcms/detail/34.1294.TN.20230516.1539.004.html, published online May 17, 2023

Manuscript received: 2023-03-15

Abstract: With the rapid advancements in edge computing and artificial intelligence, federated learning (FL) has gained momentum as a promising approach to collaborative data utilization across organizations and devices, while ensuring data privacy and information security. In order to further harness the energy efficiency of wireless networks, an integrated sensing, communication and computation (ISCC) framework has been proposed, which is anticipated to be a key enabler in the era of 6G networks. Although the advantages of pushing intelligence to edge devices are multi-fold, some challenges arise when incorporating FL into wireless networks under the umbrella of ISCC. This paper provides a comprehensive survey of FL, with special emphasis on the design and optimization of ISCC. We commence by introducing the background and fundamentals of FL and the ISCC framework. Subsequently, the aforementioned challenges are highlighted and the state of the art in potential solutions is reviewed. Finally, design guidelines are provided for the incorporation of FL and ISCC. Overall, this paper aims to contribute to the understanding of FL in the context of wireless networks, with a focus on the ISCC framework, and provide insights into addressing the challenges and optimizing the design for the integration of FL into future 6G networks.

Keywords: integrated sensing; communication and computation; federated learning; data heterogeneity; limited resources

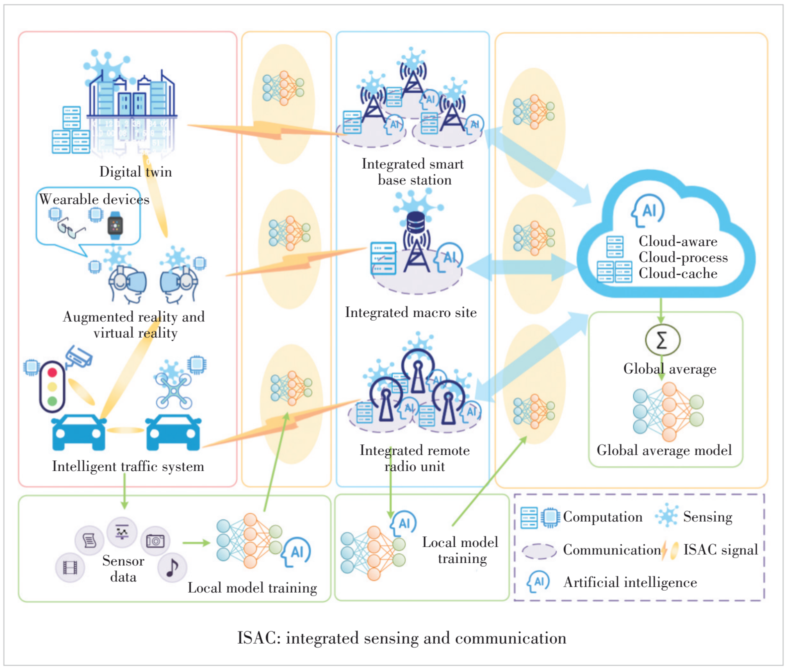
Citation (Format 1): ZHAO M K, HUANG Y S, LI X. Federated learning for 6G: a survey from perspective of integrated sensing, communication and computation [J]. ZTE Communications, 2023, 21(2): 25 – 33. DOI: 10.12142/ZTECOM.202302005

Citation (Format 2): M. K. Zhao, Y. S. Huang, and X. Li, "Federated learning for 6G: a survey from perspective of integrated sensing, communication and computation," *ZTE Communications*, vol. 21, no. 2, pp. 25 – 33, Jun. 2023. doi: 10.12142/ZTECOM.202302005.

1 Introduction

With the continuous integration and advancement of communications and the popularization and application of artificial intelligence (AI), the next-generation communication system will not only facilitate huge data rates but also enable the intelligent industry of the Internet of Things (IoT)[1]. The number of connected devices worldwide is estimated to reach 29.3 billion[2] by 2023. The entire IoT network will provide low-latency, high-precision, scalable and flexible services powered by AI and non-contact sensing techniques[3]. In conventional wireless networks, high-quality environmental data are gathered by sensing and then conveyed via data transmission links, which is finally computed in downstream tasks. These separate processes pose challenges to meeting the stringent requirements of ultra-low latency, high reliability, and high capacity in 6G networks.

An integrated network can realize closed-loop information flow and wide-area intelligent cooperation (Fig. 1). It profoundly integrates wireless sensing func-



▲Figure 1. Application scenarios of integrated communication, sensing and computation tions, including but not limited to traditional positioning, detection, imaging, and wireless transmission. It also leverages widely distributed computing power to process additional computations, thus realizing the cross-fusion of perceptual communication computing and supporting typical application scenarios such as the smart city, intelligent healthcare, and smart home in 6G systems. To enhance the ability of 6G networks to perform endogenous intellectual sensing and adapt computational power, both academia and industry have preliminarily explored integrated sensing, communication and computation (ISCC) frameworks. On the other hand, the cloud computing platform is utilized for centralized data processing and training by machine learning (ML). Nonetheless, the vast volume of data created by intelligent terminals at the network edge may require substantial communication resources. When computation workloads are distributed for multiple tasks and all data are uploaded to a cloud platform, secure data privacy becomes difficult. Therefore, sending all data to the cloud for computing and learning may be impractical.

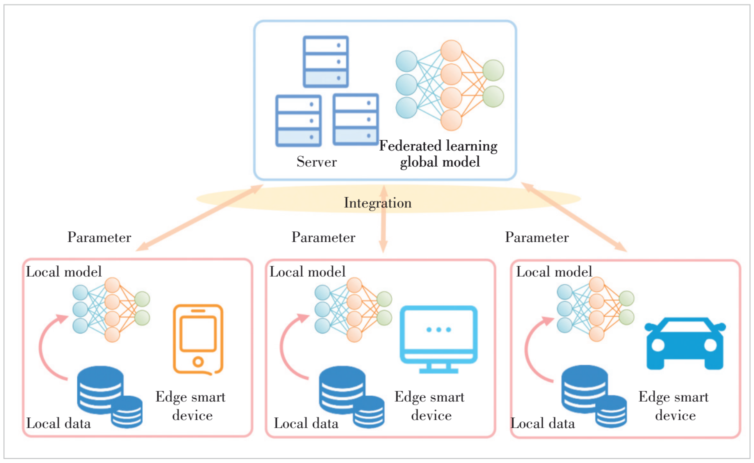
In recent years, China and the European Union have respectively introduced relevant regulations such as the Data Security Law and the General Data Protection Regulations[4], which states the regulatory requirements to ensure privacy and security while sharing data. For AI technology, federated learning (FL) was proposed for the sake of low latency and high accuracy[5]. Edge computing pushed cloud services from the network core to the edge closer to IoT devices[6]. Communication transmission delays can be efficiently decreased by intelligently combining terminal equipment, edge server and cloud center to participate in model training and data processing at the edge. Specifically, FL refers to edge intelligent sensing devices that use their computing capabilities to learn local data and obtain models based on different tasks. FL is a widely used distributed learning model, which uses wireless networks to bring a global ML model that improves computing ability and keeps data confidential to a certain extent[7].

In a typical FL training process, the central server broadcasts the global model to each edge device available. The edge device learns from the local data and obtains a local model. The regional model parameters are uploaded to the central server for aggregation to generate a new global model. This process is repeated iteratively to obtain the final global model. The federation has four leading performance indicators: latency, energy, reliability, and large-scale connectivity[8]. Because there is no need to share and transmit raw data and a cluster-like communication structure is adopted, FL is more suitable for large-scale intelligent devices and widely distributed deployment environments. In this paper, we will examine the issues faced by FL and the latest advancements in FL to investigate the future 6G network of universal computing. We will present the challenges in three categories: addressing terminal/data heterogeneity and model variances, executing FL within the constraints of universal computing resources, and bolstering privacy protection. By introducing the fundamental concepts of FL, summarizing the advantages and disadvantages of existing research, and investigating application schemes for different task scenarios, this paper discusses the research trend of FL in the future edge intelligence system. Section 2 demonstrates the basic framework of FL. In Section 3, we present techniques used for ISCC. In Section 4, we highlight several challenges when implementing FL in the ISCC framework, including participant selection, adaptive aggregation, incentive mechanism, model compression, and privacy protection. Furthermore, we review the solutions to these problems along with their advantages and disadvantages. Finally, we design guidelines for the incorporation of FL and ISCC as well as a range of typical FL applications in Section 5.

2 Framework of Federated Learning

In traditional ML systems, users are required to upload local data to cloud servers with solid computing power for centralized model training, which includes central servers and several edge devices for data collection, as shown in Fig. 2. This scenario generates energy consumption and communication delays during the data upload process. Additionally, there is also a risk of privacy disclosure for privacy-sensitive participating nodes. In response to this problem, researchers have conducted the research on distributed ML. In 2017, Google first proposed FL technology[7]. Since then, FL has received significant attention from the academic community.

In a typical training cycle of FL, the dedicated edge server initially broadcasts a global machine-learning model to participating edge devices. Subsequently, the edge devices utilize their local data to calculate their respective model updates and transfer them to the edge server for further aggregation and global model updates. The FL training process is carried out iteratively in multiple communication rounds. The dataset of N edge devices is $\{D_1, D_2, D_3, \cdots, D_N\}$. The traditional method is to upload the dataset and train the model in the central server, whereas FL coordinates the local training of many data users



▲Figure 2. Basic schematic diagram of federated learning (FL) through the network to update the parameters interactively with the global model on the server side. It cooperatively optimizes the common objective function to obtain the final ML model. For client k with dataset D_k, the loss function can be expressed as $F_k(\omega) = \frac{1}{D_k}\sum_{j \in D_k} f_j(\omega)$. $f_j(x_j, y_j, \omega)$ is the loss function of the j-th data sample related to a specific ML model, where (x_j, y_j) represents the sensing data sample, while ω is the parameter of the learning model. In each learning round t, the steps detailed below are performed.

- Sensing data collection: each edge device k is equipped with sensing capabilities and collects data samples for local model training.
- Global model broadcast: each edge device k downloads the global model parameter $\omega^{(t)}$ through wireless communication links from the central server.
- Local model training: each edge device k uses the global model $\omega_k^{(t)}$ to update its local model, where $\omega_k^{(t)}$ is the local model parameter set of client k in the round of t. Therefore, $\omega_k^{(t)} = \omega^{(t-1)} - \eta \nabla g_k(\omega^{(t-1)})$, where η indicates the learning rate. As long as the local gradient $g_k(\omega^{(t-1)})$ is obtained, $\omega_k^{(t)}$ can be calculated.
- Local model uploading: $\omega_k^{(t)}$ is sent to the server via the uplink wireless channel by using the communication mode of the client device.
- Model parameter aggregating: the local models received by the server from all devices are aggregated to obtain a new global ML model, that is, $\omega^{(t+1)} = \frac{1}{K}\sum_{k \in K}\omega_k^{(t)}$.

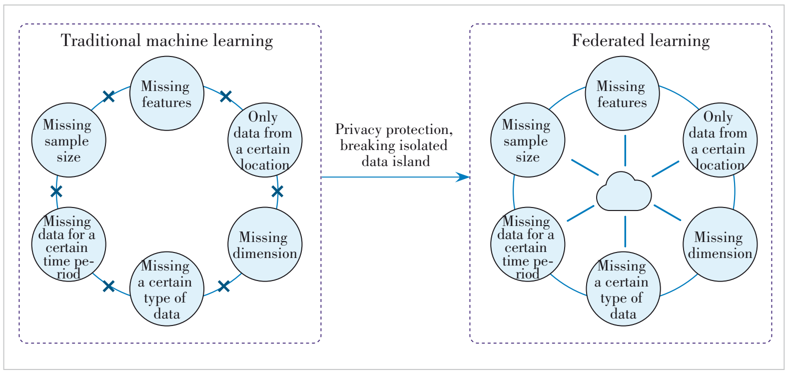
By leveraging the FL framework, the initial dataset is stored locally and trained on edge devices. This eliminates the need for sharing data with other devices or servers and further ensures that only global models are achieved via the transmission of model parameters. This advantage may address the limitations of traditional ML methods. As shown in Fig. 3[9], FL achieves the construction of unified data among multiple nodes, providing higher quality services for big data applications by increasing data sample size, data types, data features, and data dimensions, and creating value for the future development of society. In comparison to traditional methods that gather data and train models based on cloud platforms, FL may be better equipped to handle dispersed computing tasks, while simultaneously preserving the privacy of user data. Additionally, FL may help alleviate the exponential increase in cost that arises from an increased data volume. It is also believed to be more user-friendly for larger mobile terminal scales and to provide advantages for a wider distributed deployment environment. By enabling the sharing and fusion of heterogeneous device data, FL may provide powerful support for future 6G environments.

3 Integrated Sensing, Communication and Computation

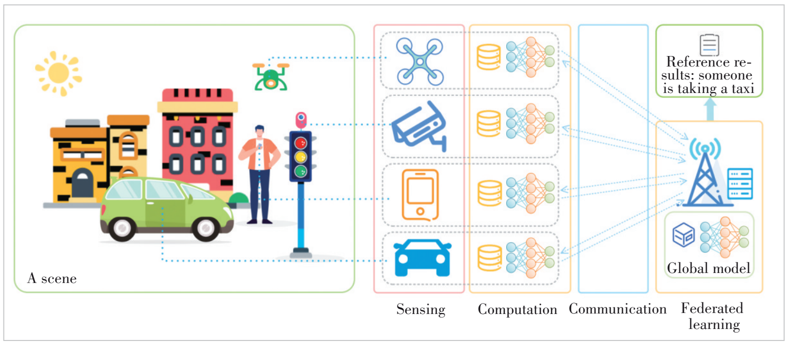
In traditional wireless networks, sensing, communication and computation are designed separately for various purposes. The isolated design principle is difficult to adapt to the strict requirements of emerging 6G applications, such as autopilot and virtual reality, which demand ultra-low latency, ultra-high reliability, and high capacity. Therefore, a new paradigm has emerged, which integrates communication and computation and comprehensively considers the application of data in downstream in a task-oriented way. As shown in Fig. 4, sensing, computation, and communication are highly coupled with FL in this new paradigm. Because of the fact that radio signals can be utilized for wireless communications and environmental sensing simultaneously, intelligent devices can analyze information about the detected target via wireless sensors in terms of range, positioning, imaging, etc. LIU et al. proposed a resource allocation approach toward ambient intelligence[10]. LI et al. introduced ISCC into over-the-air computation (AirComp) to improve spectrum efficiency and sensing performance, where function calculation from different user data is implemented by utilizing the overlay feature in wireless signal transmission in the air[11]. However, there is still a paucity of studies on the FL under the umbrella of ISCC in wireless networks.

3.1 Integrated Sensing and Communication

Integrated sensing and communication (ISAC) refers to the integration of sensing and communication into the unified de-



▲Figure 3. Comparison of traditional machine learning (ML) and federated learning (FL)[9]



▲Figure 4. Sensing, communication and computation coupled with federated learning (FL) sign of wireless networks to improve spectral efficiency and achieve mutual benefit through sensor-assisted communications and communication-assisted sensing. Compared to traditional wireless networks, ISAC can use wireless infrastructure, spectrum and power resources for simultaneous communication and sensing, which is believed to improve system performance at a lower cost. Meanwhile, the primary challenge in ISAC is the tradeoffs between performance caused by the sharing wireless resources and the contradiction between sensing and communication. These tradeoffs include information theory limitations, physical performance, propagation channel, and cross-layer indicators[12]. There are three perception tasks: detection, estimation and recognition, which are all performed based on the collected signal or data information related to the sensing object.

The integration gain can be obtained through the development of a dual-function waveform that can sense and communicate simultaneously based on shared resources. The leading methods to attain this include scheduling orthogonal or non-overlapping wireless resources (time division/frequency division/space division/code division), using separate signal waveforms, and balancing communication and sensing performance for signal waveform sharing[12]. For example, LIU et al.[13] proposed a privacy protection vertical FL scheme based on distributed ISAC for cooperative object/human motion recognition. The method uses a dedicated frequency-modulated continuous wave signal for each edge device's target sensing and data exchange. It then converts the sensing data into a low-dimensional intermediate vector and transmits it to the edge device. LI et al presented two new FL algorithms that use compression sensing to reduce the communication burden in an IoT system[14]. JEON et al. proposed a compression sensing method for FL on large-scale multi-input multi-output communication systems, which is superior to the traditional linear beamforming method[15]. It can also reduce the performance gap between FL and centralized learning through reconstruction. Based on the above, orthogonal or non-overlapping wireless resources may help to reduce functional interference, but there will be resource competition between sensing and communication. And effective management of waveform interference is necessary to separate waveforms at the same frequency. These methods can improve the efficiency of the system's spectral, hardware, and information processing efficiency, but they come with higher computational complexity.

3.2 Integrated Communication and Computation

As data volumes from edge devices rapidly increase, it becomes challenging for edge servers to receive large amounts of data from edge devices quickly through wireless links due to limited wireless communication resources. This issue may be addressed through asynchronous communication and computation resource management and AirComp. AirComp is used to integrate computation into communication, which improves communication and computation efficiency, protects user privacy, enhances user experience, and reduces delays caused by multiple access according to Ref. [16]. Compared with AirComp, communication and computing are in order in asynchronous communication and computation resource management. Optimizing resource management for asynchronous communication and computation can minimize energy consumption and delay to a certain extent, but scheduling complexity must be reduced. Compared with asynchronous communication and computation resource management, AirComp needs to consider the interference between AirComp and conventional communication applications, expand the scale of equipment, and conduct extensive performance evaluation under actual settings.

3.2.1 Asynchronous Computation and Communication

The ability of edge devices to update and upload the model status information to the edge server largely depends on their wireless channel qualities. When the edge devices operate under poor wireless channel conditions, it leads to longer model update time, which may delay the follow-up training. During model training, it is necessary to allocate wireless resources properly to improve learning performance, so that limited wireless resources can be fully utilized by collaborative asynchronous computation and communication resource management.

Various optimization algorithms can be employed to solve the problem of energy-delay allocation, or communication and computation performance can be adjusted to reduce energy consumption. On the one hand, it can be achieved by increasing bandwidth utilization. To further reduce the communication delay, ZHU et al.[17] roposed a broadband analog aggregation access scheme, which exploits the waveform superposition characteristics of the multi-access channel to achieve integrated communication and computation. The communication delay is independent of the number of edge devices connected to the access channel. On the other hand, it can also speed up the learning process to solve the problem of communication and resource constraints. For example, it can maximize the FL algorithm convergence rate under the conditions of power and energy[18–19]. However, most scenarios are characterized by coexisting sensing, communication and computation. Resource competition between communication and sensing also affects the performance of FL.

3.2.2 AirComp

Existing research mainly applies AirComp to different scenarios, or uses various optimization methods to minimize the signal mean square error[20–22] and also optimizes the model in FL by different means based on AirComp.

AMIRI et al. introduced error accumulation and how to sparsify gradient estimates based on AirComp. They proposed saving untransmitted gradient vectors in an error accumulation vector, updating the local model, and computing a new gradient vector based on this error accumulation vector in the next iteration[23]. CAO et al. proposed a "one-time" aggregation to improve communication efficiency while considering a new power control design to maximize the convergence rate[24]. The results show that the proposed power control strategy achieves a significantly faster convergence rate in FL than the fixed power control benchmark strategy. Since the error accumulation vector and gradient thinning correct the gradient computation process and make more efficient use of the bandwidth, this scheme improves the accuracy of the model based on AirComp. YANG et al. proposed a general integrated communication and computation scheme based on AirComp[25]. However, the experiment revealed that the model's accuracy would gradually decline due to the parameter aggregation mistake in AirComp caused by signal distortion.

4 Challenges in Federated Learning and Their Solutions

FL may partially solve the problem of limited computing and communication resources while preserving client privacy to a certain extent. Despite this, such technology still faces issues such as data heterogeneity, insufficient training accuracy, and low training efficiency. Various schemes are proposed to resolve these issues, as summarized in Table 1.

4.1 Participant Selection

Intelligent edge devices have limited computing capacity, making centralized data processing challenging. Thus, it is crucial to choose edge devices based on data heterogeneity and local models. However, present methods, like the time fairness scheme and throughput fairness scheme[26], may overlook the differences between learning tasks, leading to poor learning performance. Furthermore, due to the unique characteristics of edge intelligence, model updating requires wireless channel resources. Therefore, selecting the correct participants for each FL training round is vital[27].

4.1.1 Selection Based on Participants and Training Quality

Terminal equipment exhibits heterogeneity and non-independent identically distributed characteristics, resulting in substantial differences between terminal equipment models[28]. It is crucial to carefully select high-quality participants, efficiently train models, and ensure their robustness. To achieve this, different strategies are proposed. One strategy involves selecting high-quality participants. For example, LAI et al. proposed a client selection framework, which enables the identification and selection of valuable clients for training[29]. In another study, MONDAL et al. presented a distributed participant selection algorithm that minimizes the costs of energy consumption and data transmission while selecting the least number of participants with the same coverage[30]. ZHANG et al. utilized the FL framework to train lightweight neural networks that establish the relationship between context and sensor data quality[31]. Their approach leverages participants' context information to predict sensing data quality.

The other strategy is that participants' selection can be based on the quality of their local models. KATHAROPOULOS et al. proposed a power-of-choice strategy commonly used in the queue system[32]. According to their analysis, selecting the loss value as an important metric for the client can improve the convergence rate of the entire model. SATTLER et al. devised the clustered FL algorithm which divides the client into two partitions using their cosine similarity and checks partition consistency by testing the gradient norm of the client[33].

4.1.2 Selection Based on Improving Resource Management

Due to the large number of participants, the upload link may become congested, and differing participants could result in unproductive training rounds[34]. Selecting the best user cluster that aligns with limited communication and computational resources will ultimately help improve training efficiency, reduce training time, and enhance model accuracy. RIBERO et al. sug-

▼Table 1. Challenges in federated learning (FL) and their state-of-the-art solutions

Challenge	Specific Method	Advantages and Disadvantages
Participant selection	Participating clients are selected based on the heterogeneous nature of the data, quality of participants and training, and resource constraints.	Selecting participants can make full use of resources and is conducive to continuous training. However, when the data scale is too large, the overall performance cannot be guaranteed in the scenario of edge intelligence applications, and the training process needs to be optimized.
Adaptive aggregation	The best tradeoff is found between local updates and global parameter aggregation under a given resource budget to speed up the local training process.	By adapting the frequency of global aggregation, the performance of the model can be improved, and the utilization of available resources can be improved. However, the convergence of adaptive aggregation schemes currently only considers convex loss functions.
Incentive mechanism	FL requires an effective incentive mechanism for participation and balances rewards and limited communication and computing resources to improve data quality.	By quantifying data quality, the overall benefit of FL is generally improved, but due to the heterogeneity of the environment, the excitation obtained by different edge devices in FL does not match, making it difficult to balance game rewards and resource consumption.
Model compression	The transmission model is compressed to improve the communication efficiency between the server and client. Knowledge distillation exchanges model outputs, allowing edge devices to adopt larger local models.	Client-to-server parameter compression may cause convergence problems, increase computational complexity, and reduce training accuracy. Knowledge distillation alleviates the problem of independent and identical distribution of data to a certain extent, but the quality of wireless channel will affect the accuracy of model training.
Privacy protection	Privacy protection may be achieved through the inference of attacks, the encryption of data and models, and the improvement of privacy protection performance by blockchain technology.	FL may solve the privacy leakage problems caused by the model parameter sharing and multi-party communication and cooperation mechanism of FL. However, further research is needed when it comes to the security problems caused by data poisoning and the removal of traces left by participants' data in the local model, etc.

gested that only transmitting client updates with a significant amount of information during each training round reduces the transmission pressure of FL[35]. This approach decreases communication costs during the training process while ensuring model accuracy by selecting the clients participating in each update. ABDULRAHMAN et al. proposed a multi-criteria participant selection algorithm that considers participants' central processing unit, memory, energy and task completion time for FL in IoT's resource-constrained environment[36]. This algorithm maximizes the number of participants while minimizing the number of communication rounds. To ensure long-term performance, XU et al. explored FL in typical wireless networks, identifying issues related to participant selection and bandwidth allocation in long-term client energy constraints. They proposed an online optimization algorithm based on Lyapunov to address these issues[37].

4.2 Adaptive Aggregation

In FL, the model updating procedure is primarily split into two steps: local model updating at clients and global aggregation, which involves uploading model parameters to the server and aggregating them into a global model. The adaptive aggregation problem of FL specifically aims at bandwidth aggregation and model parameter aggregation[38]. With limited resources, local model updating and global aggregation are modified to accelerate convergence and improve accuracy.

GUHA et al. suggested a single-round communication federation learning system to reduce the communications between clients and servers[37]. The entire training is carried out on the edge device, and only the local model parameters are uploaded and aggregated after the movement. Based on the greedy algorithm, HADDADPOUR et al. proposed a hypercluster algorithm that trained each local model several times using the client's local data and selected the model with the minimum training loss[39]. WANG et al. proposed a control algorithm to achieve the ideal tradeoff between the local update and the global aggregation[40]. Analyzing the convergence boundary of distributed gradient descent of FL, it minimizes the training loss under a given resource budget. ZHANG et al. proposed an FL framework with adaptive local aggregation, which captured the personalized data required by the client in the global model, downloaded the global model and local model for the adaptive aggregation, and initialized the local model on each client before trained in each iteration[41].

In the same training iterations, adaptive aggregation FL reaches better performance than the synchronous aggregation of all clients. With effective utilization of computation and communication resources, it obtains lower training loss and higher model accuracy and reduces the load of edge servers.

4.3 Incentive Mechanism

The incentive mechanism quantifies the quality of data that edge devices provide to reduce energy consumption and improve model accuracy with the guarantee of data privacy and the lowest possible computation and communication costs[42]. The incentive mechanism for FL participants usually regards the edge device as the seller and provides them with training services while the server is regarded as the buyer of the data.

To improve the energy efficiency of model transmission, FENG et al. proposed a cooperative relay network-assisted parameter transmission scheme and corresponding service pricing mechanism[43], modeling the relationship between edge devices and FL servers as a Stackelberg game model[44]. SUN et al. investigated the air-ground dynamic digital twinning and joint learning and, on this basis, studied the FL incentive mechanism based on the Stackelberg game and proposed an adaptive adjustment incentive mechanism for the best user and customer selection in dynamic networks[45]. To ensure that the incentive budget is proportional to the value of the FL model and prevent the server from being forced to pay redundant rewards, RICHARDSON et al. proposed an incentive scheme based on influence to prevent the participants from receiving rewards due to redundant data[46]. Optimizing the FL incentive mechanism can effectively limit the number of the participants who falsely contribute to the work, reduce their motivation to phonily report expenses, and thus improve the overall performance of FL.

4.4 Model Compression

While the number of mobile devices rises sharply, it is challenging for mobile virtual network operators to provide low-cost and reliable access services for users due to deficient network infrastructure. The amount of uploaded data is also gradually growing in tandem with the widespread use of powerful ML on edge devices, resulting in significant bandwidth consumption and a decline in communication efficiency. Therefore, reducing the communication overhead in FL becomes an impending issue, which can be addressed by data compression, knowledge distillation, asynchronous parameter update, etc.[47].

4.4.1 Compressed Data Transmission

Compressing the transmission data is an effective measure to improve transmission efficiency. In the FL framework, model parameter compression technologies, such as network pruning, quantization and weight sharing, can be applied to reduce communication costs. Based on the uploaded information on gradient changes in the FL process, the model output value or intermediate value can also be compressed by gradient compression in different levels[48].

However, during data compression, noise is inevitable and will cause a discrepancy between the convergence result and the ideal solution, negatively impacting the effectiveness of FL. ROTHCHILD et al. proposed to reduce the number of communication rounds in FL by directly retrieving the latest gradient value without updating its position in the vector[49].

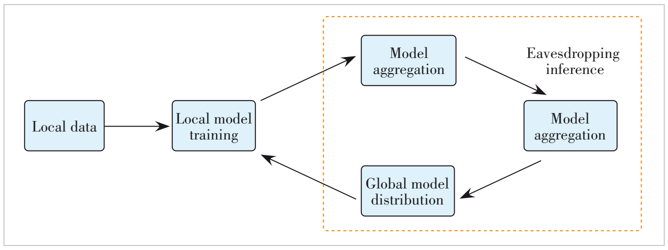
4.4.2 Knowledge Distillation

Knowledge distillation (KD) can transfer knowledge from one neural network to another by exchanging soft predictions rather than the whole model[47]. KD loss includes mild loss and local training loss. KD is employed to mine the global model knowledge[50]. In addition, LI et al. proposed the federated model distillation algorithm to train heterogeneous models in a way that protects privacy[51]. LIN et al. adopted integrated distillation to migrate the knowledge of all heterogeneous client models to the global model in each global iteration[52]. However, these algorithms ignore the further personalized needs of clients participating in training. ZHANG et al. proposed the knowledge transfer personalized FL algorithm, which parameterizes the similarity of paired clients and uses KD to transfer the personalized soft prediction knowledge to the local[53]. CHO et al. used cluster co-distillation to migrate the understanding of clients with similar data distribution to the local model[54]. DIVI et al. used KD to solve the problem of FL data heterogeneity and proposed the personalized FL algorithm, which was carried out in two stages[55]. The first stage simplies FL. In the second stage, each user selects the best teacher model from the global model of each iteration and distills it to achieve personalization.

4.5 Privacy Protection

FL is a method that is believed to effectively solve the problem of user privacy disclosure, since it does not require edge devices to upload their data. The user equipment builds a network model, generates local model parameters based on local data, and uploads them to the central server, which aggregates local model parameters to the global model, effectively protecting user privacy and easing the burden of communication bandwidth. The gradient should be protected because even if the training data stay inside the local area, attackers can still exploit these shared gradients to reverse the content of the original training data[56], exposing the training data to the public. At the same time, malicious participants or collaborators can use the intermediate information transmitted during the FL training process to launch member inference attacks[57] or data reconstruction attacks[58], exposing participants in FL to privacy disclosure threats. Fig. 5 shows potential privacy attacks in the processes of FL.

By encrypting gradient parameters, it is possible to solve the privacy disclosure issue that might arise during the process of uploading and downloading model parameters. At the same time, some researchers have focused on improving encryption efficiency[59]. FL can consistently use the blockchain consensus mechanism to establish authentic interaction in an untrusted environment. The benefit generated from the blockchain reward mechanism can also encourage knowledge sharing in FL. The combination of blockchain technology and FL can improve data privacy and achieve performance isolation[60].



▲Figure 5. Possible privacy attacks on federated learning (FL)

5 Applications and Prospects

FL is currently in use globally. For example, HART et al. used the federated averaging algorithm to predict the next word of the mobile phone keyboard input method[61]. MUHAMMAD et al. applied FL to the recommendation system[62]. Applying FL to predict the flow of urban global cellular networks can augment the data sets and improve the prediction accuracy of the model, without the problems of complexity and no real-time. Based on the FL framework, the central node collects the model updates transmitted by the edge base station for aggregation, so as to obtain a global model with good performance. The algorithm collects the data of vehicles and their tasks as input and allocates the multi-dimensional resources according to the output results of the model to meet the time-varied resource requirements and efficiently accomplish the computation tasks in the Internet of Vehicles system. Network function virtualization technology can transform traditional network hardware resources into virtual network resources. The two-way gated loop unit based on the distributed FL framework can predict virtual network function resource requirements.

Depending on its low latency and large data processing capacity, FL can also be used in the 6G era, which blends communication, sensing and computation together. For instance, due to various influences in the air[63], large-scale unmanned aerial vehicles (UAV) swarm to avoid collisions and quickly reach the destination. FL based on the wireless network can design the flight route of a UAV swarm well and better solve UAV persistent online decision-making problems by collecting sensing data from the surrounding environment[64], supporting UAV-assisted mobile edge computing with ultra-reliable and low-delay communication[65].

Despite the advancements of FL, it can be improved in several aspects, which are shown as follows.

• Various intelligent devices, including mobile phones, cameras, UAVs, depth sensors and radio sensors, produce data samples in various modalities and have a wide range of computational capacities. Researchers should put more resources to deal with the multimodal adaptive problem in FL based on ISCC.

• The applications of FL combine sensing, communication and computation together. In other words, ISCC can better conduct FL in the future. However, the resource management problems among these three aspects still need to be solved, and the hardware devices that integrate such three functions still need to be developed.

• With the development of FL, privacy attacks against FL are also expanding. It is still necessary to improve encryption efficiency based on existing encryption algorithms and continue to explore the combination of blockchain and FL to form a new edge computing paradigm with higher security.

• The number of IoT devices is increasing, while the data generated by the equipment are also expanding rapidly. In the face of scenarios that include various IoT devices[66], the FL frameworks need to couple with the access of intelligent edge devices with different task attributes. Therefore, it still requires constant exploration of practical coupling design for them and efficiency improvement of FL while ensuring the accuracy of the model.

6 Conclusions

This paper summarizes the development of FL and classifies related technologies according to the challenges that FL faces. Among these technologies, ISCC is the most significant one for its high coupling with FL. Besides, this paper introduces the research on device, data and model heterogeneity in FL and demonstrates different challenges and the existing work about FL, including participant selection, adaptive aggregation, incentive mechanism and game model, model compression, and privacy protection. In the end, the applications and prospects of FL in reality are presented.

References

[1] LETAIEF K B, SHI Y M, LU J M, et al. Edge artificial intelligence for 6G: vision, enabling technologies, and applications [J]. IEEE journal on selected areas in communications, 2022, 40(1): 5 – 36. DOI: 10.1109/JSAC.2021.3126076

[2] Cisco. Cisco annual internet report (2018 – 2023) white paper [R]. 2020

[3] DUNNE R, MORRIS T, HARPER S. A survey of ambient intelligence [J]. ACM computing surveys, 2021, 54(4): 73. DOI: 10.1145/3447242

[4] CUSTERS B, SEARS A M, DECHESNE F, et al. EU personal data protection in policy and practice [M]. Switzerland: Springer Nature, 2019

[5] ZHU G X, LIU D Z, DU Y Q, et al. Toward an intelligent edge: wireless communication meets machine learning [J]. IEEE communications magazine, 2020, 58(1): 19 – 25. DOI: 10.1109/MCOM.001.1900103

[6] YU S, CHEN X, YANG L, et al. Intelligent edge: leveraging deep imitation learning for mobile edge computation offloading [J]. IEEE wireless communications, 2020, 27(1): 92 – 99. DOI: 10.1109/MWC.001.1900232

[7] MCMAHAN H B, MOORE E, RAMAGE D, et al. Federated learning of deep networks using model averaging [EB/OL]. (2016-02-17) [2023-03-05]. https://arxiv.org/abs/1602.05629

[8] YANG Z H, CHEN M Z, WONG K K, et al. Federated learning for 6G: applications, challenges, and opportunities [J]. Engineering, 2022, 8: 33 – 41. DOI: 10.1016/j.eng.2021.12.002

[9] China Academy of Information and Communications Technology. Federal learning scenario application research report (2022) [R]. 2022

[10] LIU P X, ZHU G X, WANG S, et al. Toward ambient intelligence: federated edge learning with task-oriented sensing, computation, and communication integration [J]. IEEE journal of selected topics in signal processing, 2023, 17(1): 158 – 172. DOI: 10.1109/JSTSP.2022.3226836

[11] LI X Y, LIU F, ZHOU Z Q, et al. Integrated sensing and over-the-air computation: dual-functional MIMO beamforming design [C]//1st International Conference on 6G Networking (6GNet). IEEE, 2022. DOI: 10.1109/6GNet54646.2022.9830500

[12] LIU F, CUI Y H, MASOUROS C, et al. Integrated sensing and communications: towards dual-functional wireless networks for 6G and beyond [J]. IEEE journal on selected areas in communications, 2022, 40: (6): 1728 – 1767. DOI: 10.1109/JSAC.2022.3156632

[13] LIU P X, ZHU G X, JIANG W, et al. Vertical federated edge learning with distributed integrated sensing and communication [J]. IEEE communications letters, 2022, 26(9): 2091 – 2095. DOI: 10.1109/LCOMM.2022.3181612

[14] LI C X, LI G, VARSHNEY P K. Communication-efficient federated learning based on compressed sensing [J]. IEEE Internet of Things journal, 2021, 8(20): 15531 – 15541. DOI: 10.1109/JIOT.2021.3073112

[15] JEON Y S, AMIRI M M, LI J, et al. A compressive sensing approach for federated learning over massive MIMO communication systems [J]. IEEE transactions on wireless communications, 2021, 20(3): 1990 – 2004. DOI: 10.1109/TWC.2020.3038407

[16] JIANG T, SHI Y M. Over-the-air computation via intelligent reflecting surfaces [C]//IEEE Global Communications Conference (GLOBECOM). IEEE, 2020: 1 – 6. DOI: 10.1109/GLOBECOM38437.2019.9013643

[17] ZHU G X, WANG Y, HUANG K B. Broadband analog aggregation for low-latency federated edge learning [J]. IEEE transactions on wireless communications, 19 (1): 491 – 506. DOI: 10.1109/TWC.2019.2946245

[18] LIN F P-C, HOSSEINALIPOUR S, AZAM S S, et al. Semi-decentralized federated learning with cooperative D2D local model aggregations [J]. IEEE journal on selected areas in communications, 2021, 39(12): 3851 – 3869. DOI: 10.1109/JSAC.2021.3118344

[19] HOSSEINALIPOUR S, BRINTON C G, AGGARWAL V, et al. From federated to fog learning: distributed machine learning over heterogeneous wireless networks [J]. IEEE communications magazine, 2020, 58(12): 41 – 47. DOI: 10.1109/MCOM.001.2000410

[20] FU M, ZHOU Y, SHI Y M, et al. UAV-assisted multi-cluster over-the-air computation [J]. IEEE transactions on wireless communications, 2022, early access. DOI: 10.1109/TWC.2022.3227768

[21] LIU W C, ZANG X, LI Y H, et al. Over-the-air computation systems: optimization, analysis and scaling laws [J]. IEEE transactions on wireless communications, 2020, 19(8): 5488 – 5502. DOI: 10.1109/TWC.2020.2993703

[22] ZHU G X, XU J, HUANG K B, et al. Over-the-air computing for wireless data aggregation in massive IoT [J]. IEEE wireless communications, 2021, 28(4): 57 – 65. DOI: 10.1109/MWC.011.2000467

[23] AMIRI M M, GUNDUZ D. Federated learning over wireless fading channels [J]. IEEE transactions on wireless communications, 2020, 19 (5): 3546 – 3557. DOI: 10.1109/TWC.2020.2974748

[24] CAO X W, ZHU G X, XU J, et al. Optimized power control design for over-the-air federated edge learning [J]. IEEE journal on selected areas in communications, 2022, 40(1): 342 – 358. DOI: 10.1109/JSAC.2021.3126060

[25] YANG K, JIANG T, SHI Y M, et al. Federated learning via over-the-air computation [J]. IEEE transactions on wireless communications, 2020, 19(3): 2022 – 2035. DOI: 10.1109/TWC.2019.2961673

[26] MCMAHAN H B, MOORE E, RAMAGE D, et al. Communication-efficient learning of deep networks from decentralized data [C]//20th International Conference on Artificial Intelligence and Statistics. AISTATS, 2017: 1178 – 1187

[27] SINGH P, SINGH M K, SINGH R, et al. Federated learning: challenges, methods, and future directions [M]//Federated learning for IoT applications. Cham: Springer International Publishing, 2022: 199 – 214. DOI: 10.1007/978-3-030-85559-8_13

[28] WANG L P, WANG W, LI B. CMFL: mitigating communication overhead for federated learning [C]//39th International Conference on Distributed Computing Systems (ICDCS). IEEE, 2019: 954 – 964. DOI: 10.1109/ICDCS.2019.00099

[29] LAI F, ZHU X F, MADHYASTHA H, et al. Oort: informed participant selection for scalable federated learning [EB/OL]. (2020-10-12)[2023-03-06]. https://arxiv.org/pdf/2010.06081v1.pdf

[30] MONDAL S, MITRA S, MUKHERJEE A, et al. Participant selection algorithms for large-scale mobile crowd sensing environment [J]. Microsystem technologies, 2022, 28(12): 2641 – 2657. DOI: 10.1007/s00542-022-05271-2

[31] ZHANG W, LI Z, CHEN X. Quality-aware user recruitment based on federated learning in mobile crowd sensing [J]. Tsinghua science and technology, 2021, 26 (6): 869 – 877. DOI: 10.26599/TST.2020.9010046

[32] KATHAROPOULOS A, FLEURET F. Biased importance sampling for deep neural network training [EB/OL]. (2017-05-31) [2023-03-05]. https://arxiv.org/abs/1706.00043

[33] SATTLER F, MÜLLER K, SAMEK W. Clustered federated learning: model-

agnostic distributed multitask optimization under privacy constraints [J]. IEEE transactions on neural networks and learning systems, 2021, 32 (8): 3710 – 3722. DOI: 10.1109/TNNLS.2020.3015958
[34] WANG T, WANG P, CAI S B, et al. Mobile edge-enabled trust evaluation for the Internet of Things [J]. Information fusion, 2021, 75: 90 – 100. DOI: 10.1016/j.inffus.2021.04.007
[35] RIBERO M, VIKALO H. Communication-efficient federated learning via optimal client sampling [EB/OL]. (2020-07-30) [2023-03-07]. https://arxiv. org/abs/2007.15197
[36] ABDULRAHMAN S, TOUT H, MOURAD A, et al. FedMCCS: multicriteria client selection model for optimal IoT federated learning [J]. IEEE Internet of Things journal, 2021, 8(6): 4723 – 4735. DOI: 10.1109/JIOT.2020.3028742
[37] XU J, WANG H Q. Client selection and bandwidth allocation in wireless federated learning networks: a long-term perspective [J]. IEEE transactions on wireless communications, 2021, 20 (2): 1188 – 1200. DOI: 10.1109/TWC.2020.3031503
[38] SHI W Q, SUN Y X, HUANG X F, et al. Scheduling policies for federated learning in wireless networks: an overview [J]. ZTE communications, 2020, 18 (2): 11 – 19. DOI: 10.12142/ZTECOM.202002003
[39] HADDADPOUR F, MAHDAVI M. On the convergence of local descent methods in federated learning [EB/OL]. (2019-10-31)[2023-03-06]. https://arxiv.org/abs/1910.14425
[40] WANG S Q, TUOR T, SALONIDIS T, et al. Adaptive federated learning in resource constrained edge computing systems [J]. IEEE journal on selected areas in communications, 2019, 37(6): 1205 – 1221. DOI: 10.1109/JSAC.2019.2904348
[41] ZHANG J Q, HUA Y, WANG H, et al. FedALA: adaptive local aggregation for personalized federated learning [EB/OL]. (2022-12-02) [2023-03-06]. https://arxiv.org/abs/2212.01197
[42] WANG Z L, HU Q, LI R N, et al. Incentive mechanism design for joint resource allocation in blockchain-based federated learning [EB/OL]. (2022-02-18)[2023-03-08]. https://arxiv.org/abs/2202.10938
[43] FENG S H, NIYATO D, WANG P, et al. Joint service pricing and cooperative relay communication for federated learning [C]//International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData). IEEE, 2019. DOI: 10.1109/iThings/GreenCom/CPSCom/SmartData.2019.00148
[44] TADELIS S. Game theory: an introduction [M]. USA: Princeton University Press, 2013
[45] SUN W, XU N, WANG L, et al. Dynamic digital twin and federated learning with incentives for air-ground networks [J]. IEEE transactions on network science and engineering, 2022, 9 (1): 321 – 333. DOI: 10.1109/TNSE.2020.3048137
[46] HU P, GU H L, QI J, et al. Design of two-stage federal learning incentive mechanism under specific indicators [C]//2nd International Conference on Big Data Economy and Information Management (BDEIM). IEEE, 2021. DOI: 10.1109/BDEIM55082.2021.00103
[47] HINTON G, VINYALS O, DEAN J. Distilling the knowledge in a neural network [EB/OL]. (2015-03-09)[2023-03-05]. https://arxiv.org/abs/1503.02531
[48] AHN J-H, SIMEONE O, KANG J. Wireless federated distillation for distributed edge learning with heterogeneous data [C]//IEEE 30th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC). IEEE, 2019. DOI: 10.1109/PIMRC.2019.8904164
[49] ROTHCHILD D, PANDA A, ULLAH E, et al. Fetchsgd: cmmunication-efficient federated learning with sketching [C]//37th International Conference on Machine Learning. PMLR, 2020:8253 – 8265
[50] LEE G, JEONG M, SHIN Y, et al. Preservation of the global knowledge by not-true distillation in federated learning [EB/OL]. (2021-06-06)[2023-03-07]. https://arxiv.org/abs/2106.03097
[51] LI D L, WANG J P. FedMD: heterogenous federated learning via model distillation [EB/OL]. (2019-10-08)[2023-03-01]. https://arxiv.org/abs/1910.03581
[52] LIN T, KONG L J, STICH S U, et al. Ensemble distillation for robust model fusion in federated learning [C]//34th Conference on Neural Information Processing Systems. NeurIPS, 2020: 2351 – 2363
[53] ZHANG J, GUO S, MA X S, et al. Parameterized knowledge transfer for personalized federated learning [C]//35th Conference on Neural Information Processing Systems. NeurIPS, 2021. DOI: 10.48550/arXiv.2111.02862
[54] CHO Y J, WANG J Y, CHIRUVOLU T, et al. Personalized federated learning for heterogeneous clients with clustered knowledge transfer [EB/OL]. (2021-09-16) [2023-03-05]. https://arxiv.org/abs/2109.08119
[55] DIVI S, FARRUKH H, CELIK B. Unifying distillation with personalization in federated learning [EB/OL]. (2021-05-31) [2023-03-06]. https://arxiv. org/abs/2105.15191
[56] MU Y Q. Deep leakage from gradients [EB/OL]. (2022-12-15) [2023-03-05]. https://arxiv.org/abs/2301.02621
[57] NASR M, SHOKRI R, HOUMANSADR A. Comprehensive privacy analysis of deep learning: passive and active white-box inference attacks against centralized and federated learning [C]//IEEE Symposium on Security and Privacy (SP). IEEE, 2019. DOI: 10.1109/SP.2019.00065
[58] ZHU L G, LIU Z J, HAN S. Deep leakage from gradients [C]//33rd Conference on Neural Information Processing Systems. NeurIPS, 2019: 14774 – 14784
[59] ZHOU C Y, FU A M, YU S, et al. Privacy-preserving federated learning in fog computing [J]. IEEE Internet of Things journal, 2020, 7 (11): 10782 – 10793. DOI: 10.1109/JIOT.2020.2987958
[60] HOU D K, ZHANG J, MAN K L, et al. A systematic literature review of blockchain-based federated learning: architectures, applications and issues [C]//2nd Information Communication Technologies Conference (ICTC). IEEE, 2021: 302 – 307. DOI: 10.1109/ICTC51749.2021.9441499
[61] XU J, GLICKSBERG B S, SU C, et al. Federated learning for healthcare informatics [J]. Journal of healthcare informatics research, 2021, 5(1): 1 – 19. DOI: 10.1007/s41666-020-00082-4
[62] MUHAMMAD K, WANG Q Q, O'REILLY-MORGAN, et al. Fedfast: going beyond average for faster training of federated recommender systems [C]//26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, 2020: 1234 – 1242. DOI: 10.1145/3394486.3403176
[63] LI X X, LI Z T, OUYANG Y, et al. Using UAV to detect truth for clean data collection in sensor cloud systems [J]. ZTE communications, 2021, 19(3): 30-45. DOI: 10.12142/ZTECOM.202103005
[64] SHIRI H, PARK J, BENNIS M. Communication-efficient massive UAV online path control: federated learning meets mean-field game theory [J]. IEEE transactions on communications, 2020, 68 (11): 6840 – 6857. DOI: 10.1109/TCOMM.2020.3017281
[65] ZHANG P Y, XIE L F, XU J. Joint placement and resource allocation for UAV allocation for UAV assisted mobile edge computing networks with URLLC [J]. ZTE communications, 2020, 18(2): 49 – 56+82. DOI: 10.12142/ZTECOM.202002007
[66] ZHANG G X, DING X J, QU Z C. Space-terrestrial integrated architecture for Internet of Things [J]. ZTE communications, 2020, 18(4): 3 – 9. DOI: 10.12142/ZTECOM.202004002

Biographies

ZHAO Moke received her BE degree in electronic engineering from Beijing University of Posts and Telecommunications (BUPT), China in 2022. She is pursuing her master's degree in electronic engineering at BUPT. Her research interests include edge computing and wireless communications in 6G.

HUANG Yansong received his BE degree in electronic engineering from Beijing University of Posts and Telecommunications (BUPT), China in 2022. He is working toward his master's degree in electronic engineering at BUPT. His research interests include integrated sensing, communication and computation in 6G.

LI Xuan (xuan.li@bupt.edu.cn) is an associate professor of Beijing University of Posts and Telecommunications, China. Her research interests include distributed learning for 6G systems, edge computing, etc.