Federated Learning for Future Intelligent Wireless Networks

Edited by
Yao Sun • Chaoqun You
Gang Feng • Lei Zhang

IEEE PRESS
WILEY

Federated Learning for Future Intelligent Wireless Networks

IEEE Press
445 Hoes Lane
Piscataway, NJ 08854

Federated Learning for Future Intelligent Wireless Networks

Edited by

Yao Sun
University of Glasgow, UK

Chaoqun You
Singapore University of Technology and Design, Singapore

Gang Feng
University of Electronic Science and Technology of China, China

Lei Zhang
University of Glasgow, UK

IEEE PRESS
WILEY

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data Applied for:

[Hardback ISBN: 9781119913894]

Cover Design: Wiley
Cover Image: © Blue Planet Studio/Shutterstock

Set in 9.5/12.5pt STIXTwoText by Straive, Chennai, India

Contents

About the Editors

Yao Sun is currently a lecturer with James Watt School of Engineering, the University of Glasgow, Glasgow, United Kingdom. He has extensive research experience and has published widely in wireless networking research. He has won the IEEE Communication Society of TAOS Best Paper Award in 2019 ICC, IEEE IoT Journal Best Paper Award in 2022, and Best Paper Award in 22nd ICCT. He has been the guest editor for special issues of several international journals. He has served as TPC Chair for UCET 2021 and TPC member for a number of international flagship conferences, including ICC 2022, VTC Spring 2022, GLOBECOM 2020, and WCNC 2019. His research interests include intelligent wireless networking, semantic communications, blockchain systems, and resource management in next-generation mobile networks. He is a senior member of IEEE.

Chaoqun You is a postdoctoral research fellow at the Singapore University of Technology and Design (SUTD). She received the BS degree in communication engineering and the PhD degree in communication and information systems from the University of Electronic Science and Technology of China (UESTC) in 2013 and 2020, respectively. She was a visiting student at the University of Toronto from 2015 to 2017. Her current research interests include mobile edge computing, network virtualization, O-RAN, federated learning, and 6G.

Gang Feng received his BEng and MEng degrees in electronic engineering from the University of Electronic Science and Technology of China (UESTC) in 1986 and 1989, respectively, and the PhD degree in Information Engineering from the Chinese University of Hong Kong in 1998. At present, he is a professor with the National Key Laboratory of Wireless Communications, UESTC of China. His research interests include resource management in wireless networks, next-generation cellular networks, etc. Dr. Feng is a senior member of IEEE.

Lei Zhang is a Professor of Trustworthy Systems at the University of Glasgow. He has combined academia and industry research experience on wireless communications and networks and distributed systems for IoT, blockchain,

and autonomous systems. His 20 patents have been granted/filed in 30+ countries/regions. He published 3 books and 150+ papers in peer-reviewed journals, conferences, and edited books. He received the IEEE Internet of Things Journal Best Paper Award 2022, IEEE ComSoc TAOS Technical Committee Best Paper Award 2019, and IEEE ICEICT'21 Best Paper Award.

Preface

It has been considered one of the key missing components in the existing 5G network and is widely recognized as one of the most sought-after functions for next-generation 6G communication systems. Nowadays, there are more than 10 billion Internet-of-Things (IoT) equipment and 5 billion smartphones that are equipped with artificial intelligence (AI)-empowered computing modules such as AI chips and GPU. On the one hand, the user equipment (UE) can be potentially deployed as computing nodes to process certain emerging service tasks such as crowdsensing tasks and collaborative tasks, which paves the way for applying AI in edge networks. On the other hand, in the paradigm of machine learning (ML), the powerful computing capability on these UEs can decouple ML from acquiring, storing, and training data in data centers as conventional methods.

Federated learning (FL) has been widely acknowledged as one of the most essential enablers to bring network edge intelligence into reality, as it can enable collaborative training of ML models while enhancing individual user privacy and data security. Empowered by the growing computing capabilities of UEs, FL trains ML models locally on each device where the raw data never leaves the device. Specifically, FL uses an iterative approach that requires a number of global iterations to achieve a global model accuracy. In each global iteration, UEs take a number of local iterations up to a local model accuracy. As a result, the implementation of FL at edge networks can also decrease the costs of transmitting raw data, relieve the burden on backbone networks, reduce the latency for real-time decisions.

This book would explore recent advances in the theory and practice of FL, especially when it is applied to wireless communication systems. In detail, the book covers the following aspects:

1) principles and fundamentals of FL;
2) performance analysis of FL in wireless communication systems;

3) how future wireless networks (say 6G networks) enable FL as well as how FL frameworks/algorithms can be optimized when applying to wireless networks (6G);
4) FL applications to vertical industries and some typical communication scenarios.

Chapter 1 investigates the optimization design of FL in the edge network. First, an optimization problem is formulated to manage the trade-off between model accuracy and training cost. Second, a joint optimization algorithm is designed to optimize the model compression, sample selection, and user selection strategies, which can approach a stationary optimal solution in a computationally efficient way. Finally, the performance of the proposed optimization scheme is evaluated by numerical simulation and experiment results, which show that both the accuracy loss and the cost of FL in the edge network can be reduced significantly by employing the proposed algorithm.

Chapter 2 studies non-IID data model for FL, derives a theoretical upper bound, and redesigns the federated averaging scheme to reduce the weight difference. To further mitigate the impact of non-IID data, a data-sharing scheme is designed to jointly minimize the accuracy loss, the energy consumption, and latency with constrained resource of edge systems. Then a computation-efficient algorithm is proposed to approach the optimal solution and provide the experiment results to evaluate our proposed schemes.

Chapter 3 theoretically analyzes the performance and cost of running FL, which is imperative to deeply understand the relationship between FL performance and multiple-dimensional resources. In this chapter, we construct an analytical model to investigate the relationship between the FL model accuracy and consumed resources in FL-enabled wireless edge networks. Based on the analytical model, we explicitly quantify the model accuracy, computing resources, and communication resources. Numerical results validate the effectiveness of our theoretical modeling and analysis and demonstrate the trade-off between the communication and computing resources for achieving a certain model accuracy.

Chapter 4 proposes an efficient device association scheme for radio access network (RAN) slicing by exploiting a federated reinforcement learning framework, with the aim to improve network throughput, while guaranteeing user privacy and data security. Specially, we use deep reinforcement learning to train local models on UEs under a hybrid FL framework, where horizontally FL is employed for parameter aggregation on BS, while vertically FL is employed for access selection aggregation on the encrypted party. Numerical results show that our proposed scheme can achieve significant performance gains in terms of network throughput and communication efficiency in comparison with some known state-of-the-art solutions.

Chapter 5 proposes a deep FL algorithm that utilizes knowledge distillation and differential privacy to safeguard privacy during the data fusion process. Our approach involves adding Gaussian noise at different stages of knowledge distillation-based FL to ensure privacy protection. Our experimental results demonstrate that this strategy provides better privacy preservation while achieving high-precision IoT data fusion.

Chapter 6 presents a novel systematic beam control scheme to tackle the formulated beam management problem, which is difficult due to the nonconvex objective function. The double deep Q-network (DDQN) under a FL framework is employed to solve the above optimization problem, thereby fulfilling adaptive and intelligent beam management in mmwave networks. In the proposed beam management scheme based on federated learning (BMFL), the non-raw-data aggregation can theoretically protect user privacy while reducing handoff costs. Moreover, a data cleaning technique is used before the local model training, with the aim to further strengthen the privacy protection while improving the learning convergence speed. Simulation results demonstrate the performance gain of the proposed BMFL scheme.

Chapter 7 proposes a double-layer blockchain-based deep reinforcement federated learning (BDRFL) scheme to ensure privacy-preserved and caching-efficient D2D networks. In BDRFL, a double-layer blockchain is utilized to further enhance data security. Simulation results first verify the convergence of BDRFL-based algorithm and then demonstrate that the download latency of the BDRFL-based caching scheme can be significantly reduced under different types of attacks when compared with some existing caching policies.

Chapter 8 aims to design a dynamic scheduling policy to explore the spectrum flexibility for heterogeneous federated edge learning (FEEL) so as to facilitate the distributed intelligence in edge networks. This chapter proposes a heterogeneity-aware dynamic scheduling problem to minimize the global loss function, with consideration of straggler and limited device energy issues. By solving the formulated problem, we propose a dynamic scheduling algorithm (DISCO), to make an intelligent decision on the set and order of scheduled devices in each communication round. Theoretical analysis reveals that under certain conditions, learning performance and energy constraints can be guaranteed in the DISCO. Finally, we demonstrate the superiority of the DISCO through numerical and experimental results, respectively.

Chapter 9 discusses FedCorr, a general multistage framework to tackle heterogeneous label noise in FL, which does not make any assumptions on the noise models of local clients while still maintaining client data privacy. Both theoretical analysis and experiment results demonstrate the performance gain of this novel FL framework.

Chapter 10 provides a general overview of the analog over-the-air federated learning (AirFL) system. Specially, we illustrate the general system architecture and highlight the salient feature of AirFL that adopts analog transmissions for fast (but noisy) aggregation of intermediate parameters. Then, we establish a new convergence analysis framework that takes into account the effects of fading and interference noise. Our analysis unveils the impacts from the intrinsic properties of wireless transmissions on the convergence performance of AirFL. The theoretical findings are corroborated by extensive simulations.

Chapter 11 investigates a FEEL-based training framework to DL-based channel state information (CSI) feedback. In FEEL, each UE trains an autoencoder network locally and exchanges model parameters via the base station. Therefore, data privacy is better protected compared with centralized learning because the local CSI datasets are not required to be uploaded. Neural network parameter quantization is then introduced to the FEEL-based training framework to reduce communication overhead. The simulation results indicate that the proposed FEEL-based training framework can achieve comparable performance with centralized learning.

Chapter 12 proposes a user-centric online training strategy in which the UE can collect CSI samples in the stable area and adjust the pretrained encoder online to further improve CSI reconstruction accuracy. Moreover, the proposed online training framework is extended to the multiuser scenario to improve performance sequentially. The key idea is to adopt decentralized FL without BS participation to combine the sharing of channel knowledge among UEs, which is called crowd intelligence. Simulation results show that the decentralized FL-aided framework has higher feedback accuracy than the AE without online training.

November 2023

Yao Sun
Chaoqun You
Gang Feng
Lei Zhang

1

Federated Learning with Unreliable Transmission in Mobile Edge Computing Systems

Chenyuan Feng[1], Daquan Feng[1], Zhongyuan Zhao[2], Howard H. Yang[3], and Tony Q. S. Quek[4]

[1] Shenzhen Key Laboratory of Digital Creative Technology, The Guangdong Province Engineering Laboratory for Digital Creative Technology, The Guangdong-Hong Kong Joint Laboratory for Big Data Imaging and Communication, College of Electronics and Information Engineering, Shenzhen University, Shenzhen, Guangdong, China
[2] State Key Laboratory of Networking and Switching Technology, School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, China
[3] Zhejiang University/University of Illinois at Urbana-Champaign Institute, Zhejiang University, The College of Information Science and Electronic Engineering, Haining, Zhejiang, China
[4] Information Systems Technology and Design Pillar, Singapore University of Technology and Design, Singapore

1.1 System Model

Consider the deployment of FL in an MEC scenario, which consists of an edge access point E and multiple users $U_1, \ldots, U_M$. An edge computing server S_E is equipped with E, while a local computing unit S_m is equipped with U_m, $m = 1, \ldots, M$. As shown in Figure 1.1, the edge computing server S_E and local computing units $S_1, \ldots, S_M$ can act as the computing server and the clients, respectively, which can interact with each other via the wireless channels between E and $U_1, \ldots, U_M$.

As introduced previously, FL can be implemented by updating the local models and the global model iteratively. In particular, we focus on the tth iteration, which can be introduced as follows.

1.1.1 Local Model Training

In this phase, each user updates the local model independently based on its local collected data. Without loss of generality, we focus on a specific user U_m, the local model of U_m can be updated as follows:

$$\mathbf{w}_{t,m} = \mathbf{w}_{t-1,m} - \eta_t \nabla F(\mathbf{w}_{t-1,m}, \mathcal{D}_{t,m}), \tag{1.1}$$

Federated Learning for Future Intelligent Wireless Networks, First Edition.
Edited by Yao Sun, Chaoqun You, Gang Feng, and Lei Zhang.

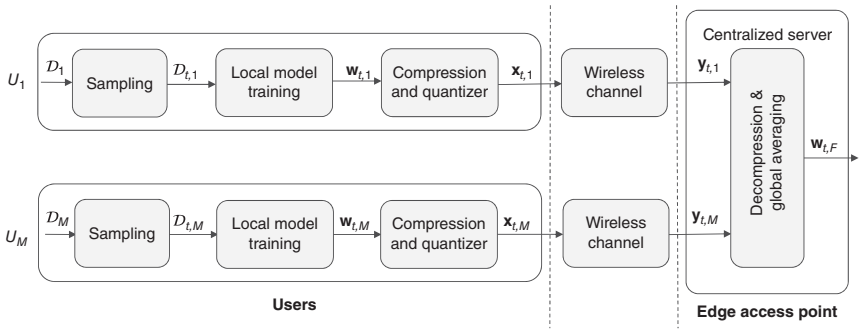

Figure 1.1 The system model of wireless FL.

where $\mathbf{w}_{t,m}$ and $\mathbf{w}_{t-1,m}$ denote the update results of U_m's local model during the tth and $(t-1)$-th iterations, respectively, $\mathcal{D}_{t,m}$ denotes the training dataset for updating $\mathbf{w}_{t,m}$, which is randomly selected from $\mathcal{D}_m$, $\mathcal{D}_{t,m} \subseteq \mathcal{D}_m$, $\mathcal{D}_m$ denotes the local dataset located at U_m, η_t is the learning rate of the tth iteration, and $\nabla F(\mathbf{w}_{t-1,m}, \mathcal{D}_{t,m})$ is the gradient of loss function $F(\mathbf{w}_{t-1,m}, \mathcal{D}_{t,m})$ with respect to $\mathbf{w}_{t-1,m}$. In this chapter, the loss function is defined as the empirical risk with respect to $\mathbf{w}_{t,m}$, which can be defined as follows:

$$F(\mathbf{w}_{t,m}, \mathcal{D}_{t,m}) = \frac{1}{N_{t,m}} \sum_{\mathbf{x} \in \mathcal{D}_{t,m}} l(\mathbf{w}_{t,m}; \mathbf{x}), \tag{1.2}$$

where $l(\mathbf{w}_{t,m}; \mathbf{x})$ denotes the loss function of the data element $\mathbf{x}$, and $N_{t,m}$ denotes the size of $\mathcal{D}_{t,m}$.

1.1.2 Update Result Feedback via the Wireless Channels

When the local model training procedure is accomplished, U_m should transmit its update result $\mathbf{w}_{t,m}$ to E via the wireless channel. In the existing works, the server randomly selects the users since it is assumed that the communications between the computing server and the clients are ideal. However, it cannot be ensured in the MEC systems due to the unreliable wireless transmission circumstances, which will cause accuracy loss of FL models. Therefore, in this chapter, only the users with high communication reliability and low model accuracy loss are scheduled to participate in each iteration of global model averaging. In particular, the scheduling status of U_m for the tth iteration of global averaging is characterized by a Boolean variable $z_{t,m}$, i.e.,

$$z_{t,m} = \begin{cases} 1, & \text{if } U_m \text{ is scheduled} \\ 0, & \text{if } U_m \text{ is not scheduled.} \end{cases}. \tag{1.3}$$

If U_m is scheduled, its update result $\mathbf{w}_{t,m}$ can be modeled as a $d \times 1$ vector, which is usually with high dimension, especially for the deep neural network models. Therefore, to improve the efficiency of update result feedback, **model sparsification** and **parameter quantization** techniques should be employed to compress $\mathbf{w}_{t,m}$. As introduced previously, $\mathbf{w}_{t,m}$ can be transformed into a sparse form via model sparsification, which can be expressed as follows:

$$\mathbf{s}_{t,m} = \mathbf{A}_{t,m}\mathbf{w}_{t,m}, \tag{1.4}$$

where $\mathbf{A}_{t,m}$ denotes a $d \times d$ sparsification matrix for $\mathbf{w}_{t,m}$.

Next, each element of $\mathbf{s}_{t,m}$ is quantized independently by employing uniform quantization. The quantization error can be approximated as an additive Gaussian noise, which is independent with $\mathbf{w}_{t,m}$. Then the quantized parameter vector can be expressed as

$$\mathbf{x}_{t,m} = \mathbf{s}_{t,m} + \mathbf{q}_{t,m} = \mathbf{A}_{t,m}\mathbf{w}_{t,m} + \mathbf{q}_{t,m}, \tag{1.5}$$

where $\mathbf{q}_{t,m}$ denotes a $d \times 1$ quantization noise vector, i.e., $\mathbf{q}_{t,m} \sim \mathcal{CN}(\mathbf{0}, \mathbf{\Omega}_{t,m})$, and $\mathbf{\Omega}_{t,m}$ denotes the covariance matrix. Due to the implementation of independent quantization, each element of $\mathbf{q}_{t,m}$ is independent with each other, i.e., $\mathbb{E}\{q_i^{t,m}(q_i^{t,m})^H\} = \omega_i^{t,m}$, $\mathbb{E}\{q_i^{t,m}(q_j^{t,m})^H\} = 0$, $i \neq j$, where $q_i^{t,m}$ and $q_j^{t,m}$ denote the ith and jth elements of $\mathbf{q}_{t,m}$, respectively. Therefore, $\mathbf{\Omega}_{t,m}$ is a diagonal matrix, which can be denoted as $\mathbf{\Omega}_{t,m} = \text{diag}\{q_1^{t,m}, \ldots, q_d^{t,m}\}$.

After model sparsification and parameter quantization, $\mathbf{x}_{t,m}$ is suitable for baseband processing and wireless transmissions. In this chapter, the flat fading model is employed to characterize the wireless channels between U_m and E. Therefore, the channel fading can be assumed to be unchanged during the transmission of $\mathbf{x}_{t,m}$. Then the observation of $\mathbf{x}_{t,m}$ at E can be expressed as

$$\mathbf{y}_{t,m} = h_{t,m}\mathbf{x}_{t,m} + \mathbf{n}_{t,m}, \tag{1.6}$$

where $h_{t,m}$ captures the flat channel fading of the wireless link between U_m and E, $\mathbf{n}_{t,m}$ denotes the additive white Gaussian noise at E, i.e., $\mathbf{n}_{t,m} \sim \mathcal{CN}(\mathbf{0}, \sigma_t^2 \mathbb{I}_L)$, $\mathbb{I}_L$ denotes a $d \times d$ identity matrix, and σ_t^2 is the power of noise.

1.1.3 Global Model Averaging

To recover the update results of local models, $\mathbf{y}_{t,m}$ should be first decompressed by E. In this chapter, the minimum mean-square error (MMSE) criterion is employed, and the decompression result can be written as

$$\overline{\mathbf{w}}_{t,m} = \arg \min_{\mathbf{w}^q \in \mathcal{C}} \|\mathbf{D}_{t,m}\mathbf{y}_{t,m} - \mathbf{w}^q\|^2, \tag{1.7}$$

where $\mathbf{D}_{t,m}$ is a $d \times d$ decompression matrix of $\mathbf{y}_{t,m}$, $\mathcal{C}$ denotes a set that consists of all the possible quantized parameter vectors, i.e., $\mathbf{w}^q \in \mathcal{C}$. Since each element of

the quantized model parameter vector can be detected individually, recalling the computational complexity of MMSE, the complexity of this detection is a linear function of the vector dimension.[1]

Then the global model can be updated by averaging the decompressed results of local models. As introduced in Konecný et al. [2016], the update result of global model can be expressed as

$$\begin{aligned}\mathbf{w}_{t,F} &= \mathbf{w}_{t-1,F} + \sum_{m=1}^{M} \frac{N_m}{N} z_{t,m} \left(\overline{\mathbf{w}}_{t,m} - \mathbf{w}_{t-1,F}\right) \\ &= \sum_{m=1}^{M} \frac{N_m}{N} \left[(1 - z_{t,m})\mathbf{w}_{t-1,F} + z_{t,m}\overline{\mathbf{w}}_{t,m}\right], \end{aligned} \tag{1.8}$$

where $\mathbf{w}_{t,F}$ and $\mathbf{w}_{t-1,F}$ denote the global model for the tth and $(t-1)$-th iterations, respectively, $z_{t,m}$ is defined by (1.3), $N = \sum_{m=1}^{M} N_m$.

After global model averaging, $\mathbf{w}_{t,F}$ is sent back to the users. Since $\mathbf{w}_{t,F}$ are transmitted via downlink transmissions, which can acquire more radio resource and higher transmit power than the local model update phase. Therefore, it can be assumed that $\mathbf{w}_{t,F}$ is received successfully by all the users. Then the local model of each user can be updated as $\mathbf{w}_{t,m} = \mathbf{w}_{t,F}$, $m = 1, \ldots, M$.

1.2 Problem Formulation

The performance of existing learning techniques is mainly determined by the accuracy of generated learning models. It is difficult to be modeled in a tractable form, and thus cannot be optimized by employing the existing resource management schemes in the MEC systems. In this section, we first derived a closed-form upper bound of model accuracy loss, which is an efficient metric to evaluate the quality of FL models. Then an optimization problem is formulated to improve the model accuracy and training efficiency of FL with limited budget of computation and communication resources.

1.2.1 Model Accuracy Loss

As introduced in Konecný et al. [2016], the objective of model training is to minimize the expected risk, which can be estimated by employing the the empirical risk given by (1.2). Therefore, the model accuracy loss, which can be defined as the

1 To implement (1.7) at edge server, it requires to feedback sparsification matrix $\mathbf{A}_{t,m}$, and this communication overhead can be reduced significantly when $\mathbf{w}_{t,m} \in \mathbb{R}^d$ is divided into multiple segments. Moreover, the detection can be done with channel estimation, each device can feedback $\mathbf{D}_{t,m}$ instead of channel state information.

Index

Federated Learning for Future Intelligent Wireless Networks, First Edition.
Edited by Yao Sun, Chaoqun You, Gang Feng, and Lei Zhang.