Multi-Server Federated Edge Learning for Low Power Consumption Wireless Resource Allocation Based on User QoE

Tianyi Zhou, Xuehua Li, Chunyu Pan, Mingyu Zhou, and Yuanyuan Yao

***Abstract*—Federated edge learning (FEL) deploys a machine learning algorithm by using devices distributed on the edge of a network, trains massive local data, uploads the local model to update the parameters after training, and performs alternate updating with global model parameters to reduce the pressure for uplink data transmission, prevent systematic time delay and ensure data security. This paper proposes that an optimal balance between time delay and energy consumption be achieved by optimizing the transmission power and bandwidth allocation based on user quality of experience (QoE) in a multi-server intelligent edge network. Given the limited computing capability of devices involved in FEL local training, the transmission power is modeled as a quasi-convex uplink power allocation (UPA) problem, and a lower energy consumption bandwidth allocation algorithm is proposed for solution-seeking. The proposed algorithm allocates appropriate power to the device by adapting the computing power and channel state of the device, thereby reducing energy consumption. As the theoretical deduction result suggests that additional bandwidth should be allocated to those devices with weak computing capabilities and poor channel conditions to realize minimal energy consumption within the restraint time. The simulation result indicates that, the maximum gain of the proposed algorithm can be optimized by 31% compared with the baseline.**

***Index Terms*—Bandwidth optimization, federated edge learning, QoE, uplink power allocation.**

I. INTRODUCTION

THE traditional machine learning (ML) algorithm usually adopts the centralized model training method [1]–[3]. However, transmitting massive data to the central server not only causes privacy leakage but also results in uplink congestion and serious transmission delay. The distributed edge learning based on distributed ML and mobile edge computing (MEC) can significantly reduce the traffic load and end-to-end time delay in communication networks by using the computing capability and datasets of massively distributed edge devices and by co-training the shared ML model [4]–[7]. Among the available learning approaches, federated edge learning (FEL) [8] shows great potential in solving the above mentioned problem.

Compared with traditional distributed machine learning, FEL has more prospects for data privacy. Firstly, the compute node has absolute control over the data and the central server cannot directly or indirectly manipulate the data on the compute node in FEL. Secondly, in the process of data transmission, compared with traditional distributed machine learning, FEL only needs to upload local model parameters without sharing local data. While protecting data privacy, it can release the pressure of uplink and greatly reduce the amount of data transmitted.

FEL supports ML in data and model training in the mobile communication system and allows multiple-edge smart devices to complete model training and parameter sharing through local data iteration [9]. The iteration process has two parts, namely, local model training and updating and global aggregation of updated parameters. Distributed local training is performed for generating the local model. The updated local model parameters are uploaded, and the global model is optimized in the central server by analyzing the local models of smart devices, then broadcasted the updated model. Unlike traditional ML, the FEL algorithm requires users to transmit local model parameters to the base station (BS) through a wireless link and imposes additional requirements for the energy consumed in training and wireless link resource allocation. Federated learning (FL) and wireless transmission have received much research attention in recent years. For instance, Zhang et al. proposed a method that ensures minimal transmission time delay to improve FL algorithm efficiency [10]. In [11], the authors used the FL algorithm for traffic estimation to maximize user data rate. To reduce time delay, [12] proposed a partially average solution and only used the updated parameters from quick-response devices for global model updating. In addition, for the purpose of reducing communication load in FL, [13] compress gradients uploaded by edge nodes to reduce the time required for communication.

However, affected by bandwidth, edge device energy consumption, and inter-cell interference, FEL faces serious challenges in wireless link data transmission. In [10]–[13], the authors ignored the effect of edge device energy consumption and inter-cell interference on the FEL training process. Moreover, no study has attempted to jointly optimize FEL wireless resources and energy consumption based on user

Manuscript received February 15, 2021; revised September 9, 2021; approved for publication by Jiming Chen, Division II Editor, October 16, 2021.

This work is supported in part by the Natural Science Foundation of Beijing Municipality under Grants (L192022), in part by the Science and technology Project of Beijing Municipal Education Commission under grant (KZ201911232046, KM202011232002)

T. Zhou, X. Li, C. Pan, M. Zhou, and Y. Yao are with the Department of Information and Communication Engineering, Beijing Informationg Science and Technology Univercity and Baicells Technologies, email: tianyi.zh@foxmail.com, {lixuehua, chunyupan}@bistu.edu.cn, Zhoumingyu@baicells.com, yyyao@bitsu.edu.cn.

Xuehua Li is the corresponding author.

Digital Object Identifier: 10.23919/JCN.2021.000040

1229-2370/21/$10.00 © 2021 KICS

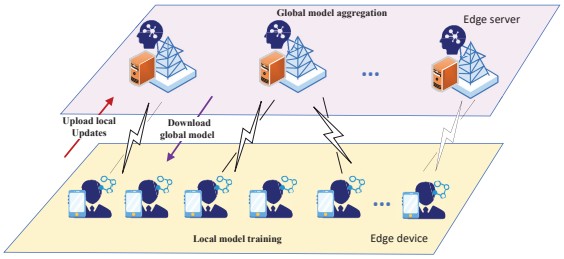


Fig. 1. Multi-unit and multi-server MEC cellular network.

quality of experience (QoE). Meanwhile, [14] selected suitable users to execute the FEL algorithm by disclosing edge device CPU cycle frequencies and transmission power to minimize energy consumption. Ref. [15] devises the Model Update Compression by Soft Clustering (MUCSC) algorithm to compress model updates transmitted between clients and the parameter server to reduce the volume of communication traffic in FL. Ref. [16] described how the computing capability and communication delay of mobile devices affect UE energy consumption, system learning time, and learning precision parameters yet only considered the application scenario of mobile cellular networks that consists of single servers. Ref. [17] obtains the user selection and uplink resource block allocation scheme by solving the optimization problem, and then deduces the optimal transmission power of each user according to the expected convergence speed of the algorithm. However, previous studies on FEL wireless transmission have also ignored multi-unit and multi-server scenarios or the effect of interference.

This paper designs a multi-unit and multi-server mobile edge cellular network for FEL wireless transmission, which maximizes network performance by implementing a strategy that jointly optimizes energy consumption and bandwidth optimization. Specifically, we consider a multi-unit MEC cellular network, with each BS equipped with an edge server for the global aggregation of the model. The distributed deployment of edge servers and intensification of BSs caters to the prospect of "smart interconnection" of all things in the future 6G system.

In this work, multi-unit and multi-server MEC networks are considered for the first time in FEL wireless transmission. Based on multi-user transmission and by considering inter-cell interference, the proposed problem is both complex and non-convex [18]. Therefore, it's hard to find a solution directly. In this case, we propose a low power consumption resource allocation strategy to reduce the complexity. The innovations of this paper are as follows:

1) Compared with single MEC server systems, each user in the designed strategy can select the closest server from the multi-server network for parameter uploading to reduce the energy consumption for FEL parameter transmission. Moreover, the coordination and allocation of resources with multiple servers can alleviate inter-cell interference, and these servers fight over wireless resources between a user and other neighboring users, thereby improving network benefits when multiple users request for FEL parameter uploading simultaneously.

2) We use task completion time and device energy consumption to quantify the QoE and to model the transmission efficiency of each user as the weighted sum of task completion time and device energy consumption optimization. Based on user QoE, the optimization problem is modeled as an uplink power allocation (UPA) to optimize the uplink transmission power of users. This paper then considers the problem of minimizing edge device energy consumption, proposes a low power consumption bandwidth allocation (BA) strategy, and theoretically deduces the convergence form of the optimal strategy for minimizing energy consumption.

3) Given the computing capability and energy consumption of edge devices involved in FEL local model training, this paper proposes a low complexity UPA algorithm for solution, which reduces the iteration times and computing complexity of the FEL.

The rest of this paper is organized as follows. Section II describes the system model. Section III proposes the uplink power allocation optimization problem for FEL wireless transmission. Section IV discusses the power consumption BA strategy. Section V provides the simulation results. Section VI summarizes the paper.

II. System Model

We consider a multi-unit and multi-server FEL system with an edge server for each BS to achieve a convergence updating of the global model. The system model is shown in Figure 1. The FEL iteration process can be divided into several steps. First, after the user utilizes local data for model training, the local model parameters transmitted through wireless links are used to update the global model. Second, the server distributes the converged and updated global model to replace the original model as shown in Figure 2. Each iteration is called a round of communication. The sets of edge devices (user) and edge servers in the FEL system are expressed as $U = \{1, 2, \cdots, u\}$ and $S = \{1, 2, \cdots, s\}$. The edge servers are assumed to obtain the model size, multi-user channel gain, local computing capability, and others through feedback. These servers use such information to determine the uplink power allocation and low power consumption bandwidth allocation strategy for each round of communication. The modeling of user computing tasks and parameter uploading is described below. The key symbols used in this paper are listed in Table I.

A. Computing Task Model

We use T_u to represent the task of user $u \in U$ that utilizes local data for model training, $\langle c_u, d_u \rangle$ to represent the computing and data amounts of the task of user u, $c_u\,[cycles]$ represents the amount of computing resources required to complete the model training, and $d_u\,[bits]$ represents the number of data consumed for uploading the parameters from edge devices to edge servers. The values of c_u and d_u can be obtained by analyzing the task execution status [19]. $f_u^l > 0$

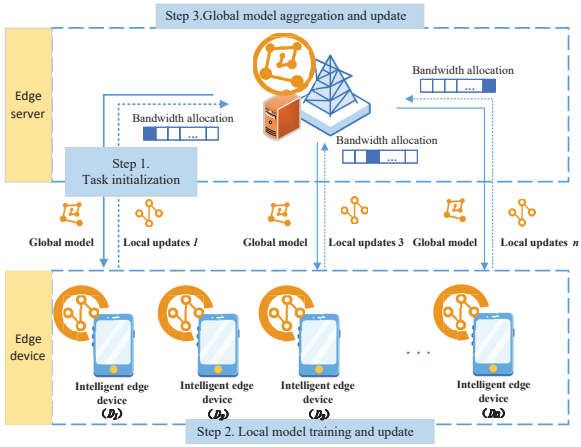


Fig. 2. FEL iterative process in a single server.

represents the local computing capability of user u with the unit of CPU cycle/s. The local model training time for user u is $t_u^l = c_u/(f_u^l)$.

B. Training Model

We design a general FEL model[17]. Each user u collects an input matrix $I_u = [i_{u1}, \cdots, i_{un}]$, where n represents the number of samples collected by each user u, i_{un} is the input vector of FL algorithm. The size of i_{un} depends on the specific FEL task. For example, the network can execute an FEL algorithm to sense the wireless environment and generate a holistic radio environment mapping, each user will collect the data related to the wireless environment for training an FL model[21]. Output of local edge device index by $o_u = [o_{u1}, \cdots, o_{un}]$. We capture the input parameters of local model training I_u and output parameter o_u by w_u, which determines the local model of each user u. For example, in the task of classification using BP algorithm, $i_{un}^T w_u$ indicates classified output, w_u is the weight vector that determines the BP algorithm.After the local model training, user u uploads w_u to the edge server to participate in the global model aggregation. Using the BP algorithm to perform local training for the neural network model on a GPU is considered [21]. The experiment in [22] reveals that the energy consumption of GPU only depends on the complexity and model parameter size (or equivalent dimensions) of the BP algorithm. Given that all edge devices use the BP algorithm to train the same models with data amount d_u, all edge devices used for local training are assumed to be the same and expressed as E_u^l. We assume that the model training of the equipment is accurate and the loss is within the acceptable range.

C. Parameter Uploading Task Model

We consider OFDMA as the mobile edge computing cellular network for the multi-address access solution in the uplink [23], and we divide the working frequency band B into N equal sub-frequency bands with a size of $W = B/N$[Hz]. $\mathcal{N} = \{1, \cdots, N\}$ represents the available sub-frequency bands of each BS. To ensure the orthogonality of uplink transmission between users related to the same BS, each user is allocated with one sub-frequency band. The N sub-frequency bands divided from each BS can serve up to N users simultaneously. We define the regulation parameter $x_{us}^j, u \in U, s \in S, j \in \mathcal{N}$ of the uplink sub-frequency band. When $x_{us}^j = 1$, the computing task T_u of user u is uploaded to the BS through channel j. Otherwise, $x_{us}^j = 0$. Moreover, we define $U_s = \left\{u \in U | \sum_{j \in \mathcal{N}} x_{us}^j = 1\right\}$ as the set of users who are uploading parameters to server s.

The delay generated in each round of communication includes 1) the time for edge devices to use local data for model training, t_u^l[s], and 2) the time for edge devices to upload parameters to edge servers through an uplink, t_u[s]. The time restraint for local model training and model parameter uploading is

$$t_u^l + t_u \leq T, \forall u \in U, \tag{1}$$

where T is the maximal time constraint. The computing capability heterogeneity of edge devices is measured by the difference in $\{t\&_u^l\}$ values. The computing capability heterogeneity of edge devices is measured by the difference in t_u^l values. $\mathcal{T} = \{t_u | 0 < t_u \leq T_u, u \in U\}$ represents the time set of users u uploading parameter d_u to edge servers. The value of t_u is limited by the maximal transmission time $T_u = T - t_u^l$.

We assume that each user and BS have a single antenna for uplink transmissions. h_{us}^j is defined as the uplink gain transmission between user u and edge server s through sub-channel j and represents the effect of path loss, shadow, and antenna gain. The association duration between the user and server is usually much longer than the small-scale decline duration. We assume that the effect of quick decline during the association period is uniform [24]. $\mathcal{P} = \{p_u | 0 < p_u \leq P_u, u \in U\}$represents user transmission power, and p_u[W] represents the transmission power consumed by user u in the uploading parameter d_u to the server. This value is limited by maximal transmission power P_u. If $u \notin U$, then $p_u = 0$. Although users send tasks to the same edge server through different sub-frequency bands, the transmission is still affected by inter-cell interference. Under this condition, the signal to interference plus noise ratio (SINR) of user u uploading parameters to the server s through sub-frequency band j can be represented as

$$\gamma_{us}^j = \frac{p_u h_{us}^j}{\sum_{r \in S} \sum_{k \in U_r} x_{kr}^j p_k h_{ks}^j + N_0}, \forall u \in U, s \in S, j \in \mathcal{N}, \tag{2}$$

where N_0 is the noise variance, and the first term of the denominator represents the total interference of all users related to other edge servers s on the same sub-frequency band j. Given that the transmission of one user is always on one sub-frequency band, the maximal transmission rate [bits/s] of the user u to server s is given by

$$R_{us}(\chi, \mathcal{P}) = W log_2(1 + \gamma_{us}). \tag{3}$$

where χ is the uploading decision. Given uploading decision χ and transmission power p_u, the time for uploading the

parameter of user u is given by

$$t_u = \sum_{s \in S} x_{us}\left(\frac{d_u}{R_{us}(\chi, \mathcal{P})}\right), \forall u \in U. \tag{4}$$

The uploading energy consumption, E_u [J], of user u can be calculated as $E_u = \frac{p_u t_u}{\xi_u}, \forall u \in U$, where ξ_u is the power amplifier efficiency of user u. We assume that $\xi_u = 1, \forall u \in U$.Therefore, the uploading energy consumption of user u can be simplified as

$$E_u = p_u t_u = p_u d_u \sum_{s \in S} \frac{x_{us}}{R_{us}(\chi, \mathcal{P})}, \forall u \in U. \tag{5}$$

In the FEL system, user QoE is mainly reflected by task completion time and energy consumption. In the scenarios considered in this paper, compared with maximal task completion time and maximal energy restraint, the relative optimization of task completion time and energy consumption is represented by $\frac{T-t_u}{T}$ and $\frac{E-E_u}{E}$. Therefore, we can define user uploading utility as

$$J_u = \left(\beta_u^t \frac{T - t_u}{T} + \beta_e^t \frac{E - E_u}{E}\right) \sum_{s \in S} x_{us}, \forall u \in U. \tag{6}$$

where $\beta_t^u, \beta_e^u \in [0, 1]$ and $\beta_u^t + \beta_u^e = 1, \forall u \in U$, which represent the preferences of user u in task completion time and energy consumption, respectively. For example, for user u, a short battery life can increase β_u^e and reduce β_u^t, that is, energy consumption is reduced by extending task completion time. In practical operations, cellphone users can set β_u^e through different power-saving modes. For instance, under the super power-saving mode, $\beta_u^e = 1$, and under the maximum performance mode, $\beta_u^e = 0$. These users can also set the parameters based on the battery levels of their devices. E represents the maximal restraint for energy consumption and is determined by the actual condition of the edge device.

D. Problem Formulation

For a given uploading strategy χ and uplink power allocation $\mathcal{P}$, we define the system utility as follows as the weighted sum of the uploading efficiency of all users:

$$J(\mathcal{P}) = \sum_{u \in U} \lambda_u J_u, \tag{7}$$

where J_u is defined in (6), and $\lambda_u \in (0, 1]$ defines the preference of the edge server in user $u, \forall u \in U$. This parameter also determines the handling priority of different edge devices. For example, based on the obtained edge device information, the devices with enough battery levels and more updating data should be prioritized with high value of λ_u. We now use a maximal system utility problem to represent power allocation as

$$\max_{\mathcal{P}, W, x_{us}^j} \; J(\mathcal{P}) \tag{8}$$

$$s.t. \quad \sum_{u \in U} x_{us}^j \le 1, \forall u \in U, s \in S, \tag{8a}$$

$$W > 0, \tag{8b}$$

Table I

Summary of Key Notations

Notation	Description
$\mathcal{U}$	Set of u edge of devices
$\mathcal{S}$	Set of s MEC servers/BSs
T_u	Training task of device u when using local data for its model
d_u	Input data of training task T_u
c_u	Workload of training task T_u
f_u^l	Local computing capability of device u
E_u^l	Energy consumption of device u when training its model locally
B	Uplink system bandwidth
$\mathcal{N}$	Set of N orthogonal sub-bands
x_{us}^j	Task uploading indicator, $\forall u \in \mathcal{U}, s \in S, j \in \mathcal{N}$
t_u^l	Local training time of model training task T_u
t_u	Transmission time of local updates to the MEC server
T	Maximum total time
T_u	Maximum transmission time of updates
h_{us}^j	Uplink channel gain between device u and MEC server s on sub-band j
p_u	Transmission power of device u
P_u	Maximum transmission power of device u
γ_{us}^j	SINR from device u to MEC server s on sub-band j
R_{us}	Uplink data rate from device u to MEC server s
χ	Updates uploading policy
E_u	Energy consumption of device u when uploading its updates
J_u	Uploading utility of device u
β_u^t	Device u's preference on task completion time
β_u^e	Device u's preference on energy consumption
λ_u	MEC server s's preference towards device u

$$0 < p_u \le P_u, \forall u \in U, \tag{8c}$$

$$0 < E_u < E, \forall u \in U, \tag{8d}$$

$$0 \le t_u \le T_u, \forall u \in U. \tag{8e}$$

The constraints in the above formulation can be explained as follows. (8a) implies that each sub-frequency band of each edge server serves one user at most. (8b) stipulates the system bandwidth. (8c) to (8e) specify the maximal transmission power, uploading energy consumption, and transmission time for each user, respectively. Given the limited computing capability and energy consumption of the involved edge devices, we aim to formulate a feasible low-complexity algorithm to the abovementioned problem.

III. Low-Complexity Power Allocation Optimization Algorithm

By exploiting the structure of the objective function and constraints in (8), we design a low-complexity algorithm to optimize transmission power allocation. Given a feasible task uploading decision χ that meets restraint (8a), we use the J_u expression in (6) to rewrite the target function in (8) as

$$J(\mathcal{P}) = \sum_{s \in S} \sum_{u \in U_s} \lambda_u (\beta_u^t + \beta_u^e) - V(\mathcal{P}), \tag{9}$$

where

$$V(\mathcal{P}) = \sum_{s \in S} \sum_{u \in U_s} \lambda_u \left(\frac{\beta_u^t t_u}{T} + \frac{\beta_u^e E_u}{E}\right). \tag{10}$$

The right side of (9) is constant for a specific uploading decision, and $V(\mathcal{P})$ can be seen as the total uploading cost of all users who have uploaded parameters. Therefore, we can redefine (8) as follows as a problem of minimizing total uploading cost:

$$\min_{\mathcal{P}} \quad V(\mathcal{P}) \tag{11}$$

$$s.t. \quad (8b)(8c)(8d) \tag{11a}$$

Moreover, from (10), (4), and (5), we obtain

$$V(\chi, \mathcal{P}) = \sum_{s\in S}\sum_{u\in U_s} \frac{\phi_u + \psi_u p_u}{log_2(1+\gamma_{us})}, \tag{12}$$

where $\phi_u = \frac{\lambda_u \beta_u^t d_u}{T_u W}, \psi_u = \frac{\lambda_u \beta_u^e d_u}{EW}$.Therefore, we define (12) as the target function of the UPA problem. Specifically, the UPA problem can be represented as

$$\min_{\mathcal{P}} \quad \sum_{s\in S}\sum_{u\in U} \frac{\phi_u + \psi_u p_u}{log_2(1+\gamma_{us})}, \tag{13}$$

$$s.t. \quad 0 < p_u \leq P_u, \forall u \in U. \tag{13a}$$

The in-cell interference $I_{us}^j = \sum_{w\in S}\sum_{k\in U_w} x_{ks}^j p_k h_{ks}^j$ of the uplink SINR γ_{us}^j of user $u \in U_s$ depends on the transmission power of other users related to other BSs on the same sub-frequency band as the cell. Therefore, problem (13) remains a non-convex problem whose optimal solution cannot be easily found. To facilitate solution seeking, we need to find the approximate value of I_{us}^j to find the solution to γ_{us}^j and therefore divide problem (13) into several sub-problems. The optimal uplink power allocation P^* obtained through solution-seeking remains the optimal value for the solution seeking of (13).

Assume that the uplink power allocation for each BS $s \in S$ is relatively independent, that is, users have no mutual collaboration or do not inform one another about their uplink transmission power between edge servers. In this case, the upper bound of I_{us}^j is

$$\tilde{I}_{us}^j \triangleq \sum_{w\in S}\sum_{k\in U_w} x_{ks}^j p_k h_{ks}^j, \forall u \in U, s \in S, j \in \mathcal{N}. \tag{14}$$

We regard $\tilde{I}_{us}^j$ as the approximate value of I_{us}^j. Given that the FEL system only selects partial users for parameter uploading for each round of communication, the value of I_{us}^j is very small, that is, a small error of $\tilde{I}_{us}^j$ will not lead to a huge difference in γ_{us}^j. By using $\tilde{I}_{us}^j$ to replace I_{us}^j, we can obtain the approximate uplink SINR value that user u uploads to edge server s through channel j as follows:

$$\tilde{\gamma}_{us}^j = \frac{p_u h_{us}^j}{\tilde{I}_{us}^j + N_0}, \forall u \in U, s \in S, j \in \mathcal{N}. \tag{15}$$

Let $\vartheta_{us} = \frac{\sum_{j\in\mathcal{N}} h_{us}^j}{\tilde{I}_{us}^j + N_0}, \Gamma_s(p_u) = \frac{\phi_u + \psi_u p_u}{log_2(1+\vartheta_{us} p_u)}$The target function in (13a) can be approximated as $\sum_{s\in S}\sum_{u\in U}\Gamma_s(p_u)$. The target function and restraint corresponding to the transmission power of each user are independent of each other. Therefore, the UPA problem described in (20) can be approximated as an optimization of the uploading power of each user $u, u \in U, s \in S$,which can be expressed as

$$\min \quad \sum_{u\in U_s} \Gamma_s(p_u), \tag{16}$$

$$s.t. \quad 0 < p_u \leq P_u, \forall u \in U. \tag{16a}$$

Problem (16) remains a non-convex problem because the p_u-related second-order derivative $\Gamma_s^{''}(p_u)$ of the target function does not meet the requirement of being constantly larger than 0. However, we can use the quasi-convex optimization technique to solve the problem (16) based on the following lemma:

a) Lemma 1: The definition field defined by $\Gamma_s(p_u)$ in (16a) is strictly quasi-convex.

Proof: See Appendix A.

Quasi-convex problems can be usually solved by dichotomy. Specifically, dichotomy finds the solution to a convex feasibility problem [25] in each round of iteration. However, the internal cutting plane method commonly used to solve convex feasibility problems requires $\mathcal{O}(n^2/\epsilon^2)$ iterations, where n is the number of problem dimensions. We also propose a method for further reducing the complexity of dichotomy.

Note that the quasi-convex function achieves the local optimum at the progressive decline point of the first-order derivative, and any local optimum of a strictly quasi-convex function is the global optimum [26]. Therefore, based on Lemma 1, we can determine that the optimal solution p_u^* of problem (16) is at the restraint bound, that is, $p_u^* = P_u$ or $\Gamma_s^{'}(p_u^*) = 0$.When equation (17) is satisfied, we can verify that $\Gamma_s^{'}(p_u^*) = 0$.

$$\Omega_s(p_u) = \psi_u log_2(1+\vartheta_{us} p_u) - \frac{\vartheta_{us}(\phi_u + \psi_u p_u)}{(1+\vartheta_{us} p_u) ln2} = 0. \tag{17}$$

We can conclude that $\Omega_s^{'}(p_u) = \frac{\vartheta_{us}^2(\phi_u+\psi_u p_u)}{(1+\vartheta_{us} p_u)^2 ln2} > 0$ and $\Omega_s(0) = -\frac{\vartheta_{us}\phi_u}{ln2} < 0$,which suggests that $\Omega_s(p_u)$ is a monotonically increasing function that is negative at the starting point $p_u = 0$. Therefore, we can design a low-complexity dichotomy to evaluate $\Omega_s(p_u)$ in each iteration instead of finding a solution to a convex feasibility problem so as to obtain the optimal solution p_u^* as shown in Algorithm 1.

In Algorithm 1, if $\Omega_s(p_u) > 0$, then the algorithm will terminate after $[log_2(P_u/\xi)]$ iterations. ξ is the convergence threshold in line 14. The time complexity of this algorithm is $O(\log_2(n))$. $P^* = \{p_u^*, u \in U\}$ represents the power allocation optimization solution for a given task uploading strategy.

IV. Low Power Consumption Bandwidth Allocation Strategy

In the previous section, based on user QoE and given task uploading solution ξ, we obtain the power allocation optimization solution $\mathcal{P}^* = \{p_u^*, u \in U\}$. To further reduce the parameter uploading energy consumption of the FEL system, we develop a low power consumption BA strategy based on the above mentioned solution.

We consider the BA problem for edge devices that satisfy the time restraint. The target of solving the BA problem is to minimize the total energy consumption, that is, $\sum_{u\in U_s}(E_u^l + E_u)$. Given that the energy consumption E_u^l for the local

Algorithm 1: Bisection Method for Uplink Allocation

Calculate $\Omega_s(P_u)$ using (17)
if $\Omega_s(P_u) \leq 0$ **then**
  $p_u^* = P_u$
else
  Set optimality tolerance $\xi > 0$
  Initialize $p_u' = 0$ and $p_u'' = P_u$
  repeat
    Set $p_u^* = (p_u' + p_u'')/2$
    if $\Omega_s(p_u^*) \leq 0$ **then**
      Set $p_u' = p_u^*$
    else
      Set $p_u'' = p_u^*$
    end if
  until $p_u'' - p_u' \leq \xi$
  Set $p_u^* = (p_u' + p_u'')/2$
end if

model training of all edge devices is equal, this problem can be transformed into minimizing uploading energy, that is,

$$\min_{\delta_u, t_u} \quad E_u \tag{18}$$

$$s.t. \quad \sum_{u \in U_s} \delta_u = 1, 0 \leq \delta_u \leq 1 \tag{18a}$$

$$0 \leq t_u \leq T_u, \forall u \in U, \tag{18b}$$

where δ_u represents the bandwidth allocation rate and $E_u^{up} = \delta_u B p_u t_u = \sum_{u \in U} \frac{\delta_u B t_u (\tilde{I}_{us}^j + N_0)}{h_{us}^j} (2^{\frac{d_u}{\delta_u B t_u}} - 1)$.Constraint (18a) means that the bandwidth sum allocated to edge devices uploading through the same frequency band does not exceed the total bandwidth, whereas constraint (18b) means that all devices involved in the uploading satisfy the time restraint. The optimal bandwidth allocation rate δ_u for edge devices can be obtained by seeking a solution to problem (18).

a) Lemma 2: The target function of problem (18) is a non-increasing function related to t_u and δ_u, $\forall u \in U$.

The validity of Lemma 2 can be easily proven by finding a solution to the target function. Based on this lemma, the optimal solution to problem (18) can be obtained by maximizing the transmission time of each device within the time restraint, that is, $t_u^* = T_u, u \in U$. and the values of t_u^* and bandwidth allocation ratio δ_u are independent of each other. Therefore, the obtained optimal BA strategy is as follows.

b) Theorem 1: The optimal BA strategy can be expressed as

$$\delta_u^* = \frac{d_u ln2}{BT_u[1 + \mathcal{W}(\frac{h_{us}^j v^* - BT_u(\tilde{I}_{us}^j + N_0)}{BT_u(\tilde{I}_{us}^j + N_0)e})]}, \forall u \in U, t_u^* = T_u, \tag{19}$$

where $\mathcal{W}(\cdot)$ is the Lambert W function, v^* is the Lagrange multiplier, and e is the Euler number.
Proof: See Appendix B.

To easily find a solution, we propose the following corollary:

c) Corollary 1: δ_u^* is a non-increasing function related to T_u and h_{us}^j.
Proof: See Appendix C.

Corollary 1 shows that edge devices with a relatively weak computing capability, that is, relatively small T_u, will limit the synchronous updating of model parameters. Therefore, additional bandwidth should be allocated to these edge devices to minimize energy consumption. Specifically, those devices with weak computing capability can complete the model parameter uploading and reduce transmission power within the transmission restraint time by receiving additional bandwidth allocation.

Additional bandwidth should also be allocated to those devices with weak channels. The problem of poor channels can be solved by either improving transmission power or increasing bandwidth. To achieve the target of minimizing energy consumption, increasing bandwidth is the optimal solution.

V. Simulation Results

The performance of the proposed system with the uplink power allocation optimization and BA strategies is evaluated based on the simulation results. The multi-server edge cellular network considered in this paper is closer to the actual scenario. However, we can adjust S to let the optimization problem presented is suitable for single server and multiple server scenarios. Unless otherwise specified, the simulation parameters are set as follows. We consider a dense heterogeneous network environment using MEC, which consists of $S = 7$ intelligent edge units. Each unit includes one BS and 50 small base stations (SBS). The coverage radius of the BS is 500 meters, and that of SBS is 5 meters. It is assumed that each small base station serves only one person. The MEC server is deployed near BS, furthermore, the edge devices obey the uniform distribution and choose the nearest base station for communication. The duration of local model training, $\{t_u^l\}$, is evenly distributed within (0, 10] ms. The channel bandwidth is B=1MHz, the uplink gain, $\{h_{us}^j\}$, of the sub-channel j between the edge device u and BS observes Rayleigh fading, the average path loss is 10^{-5}, the Gaussian noise variance is $N_0 = 10^{-8}$, the maximal transmission power of edge device u is P_u =10W, and the model size is set to $d_u = 10^4$ bits to facilitate learning. For local model training, we assign MNIST dataset to each user for classification. We build a CNN model with 6 convolutional layers, $2 * 2$ max pooling layers, a fully-connected layer and a softmax output layer.

a) **The tradeoff performance of the proposed algorithm with time and energy:** Given that only some edge devices can upload local model parameters simultaneously, each device is allocated with the same bandwidth. Based on user QoE, 50 different values are randomly set within the range of (0, 1) for the task completion time preference parameter β_u^t and energy preference parameter β_u^e of user u. Figure 3 describes the tradeoff between transmission time and transmission energy. The simulation result is shown as follows. When the uploading

Table II

Simulation Parameter List

Parameter	Numerical Value
Uplink system bandwidth B	1 MHz
Number of MEC servers/BSs S	7
Number of edge of devices U	50
Gaussian noise variance N_0	10^{-8}
Uplink channel gain h_{us}^j	10^{-5}
Model size d_u	10^4bits
The maximal transmission power of edge device P_u	10 W
The duration of local model training t_u^l	0-10 ms

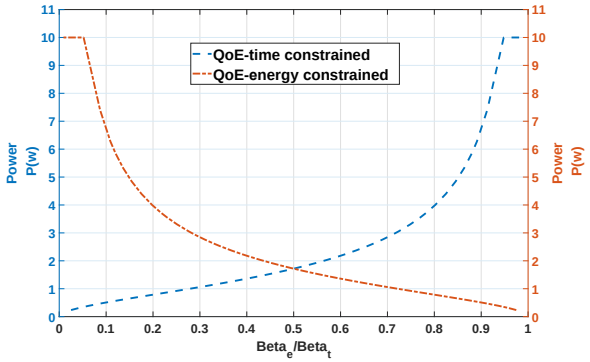


Fig. 3. Relationship between energy consumption/transmission time preference parameters and uploading power in the FEL system.

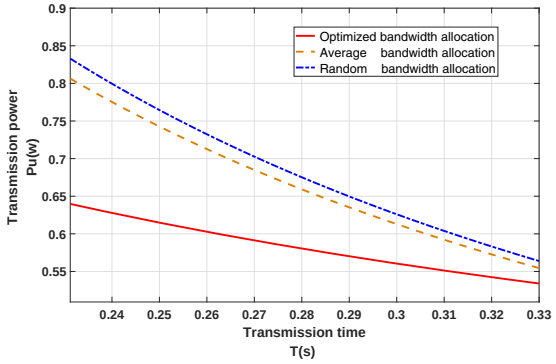


Fig. 4. Relationship between the uploading power P_u and transmission time T of edge devices.

time of model parameters is limited, a larger time preference parameter value β_u^t can be set, and the user transmission power will increase along with β_u^t. At this time, the energy preference parameter is relatively small, while the energy consumption is large. By contrast, when the energy consumed for model parameter uploading is limited, a larger energy consumption preference parameter β_u^e can be set. User transmission power decreases along with an increasing β_u^e. At this time, the task completion time preference parameter is relatively small, and the energy consumption is also small. The simulation result verifies the effectiveness of optimizing the uplink power and uploading time based on user QoE.

b) **The optimal performance of the proposed algorithm:** We then compare the practical performance of the proposed

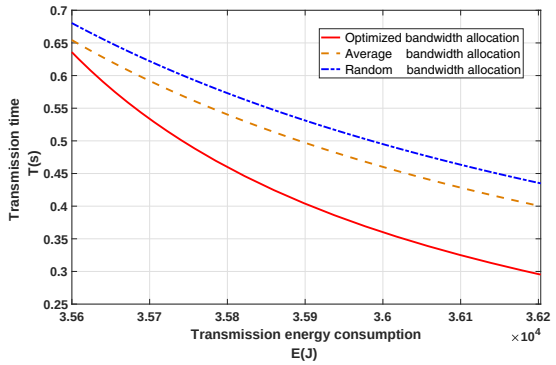


Fig. 5. Relationship curve between uploading energy consumption E_u and transmission time T.

optimal strategy with the baselines of average and random bandwidth allocations. In the average bandwidth allocation strategy, the uplink bandwidth is evenly distributed across all edge devices involved in uploading, with each device having the same uplink bandwidth. In the random bandwidth allocation strategy, the uplink bandwidth is distributed across all edge devices involved in uploading at random proportions. Based on Corollary 1, we verify the effectiveness of the proposed optimal strategy from maximal transmission time and channels.

i. The relationship curve between the uploading power P_u and transmission time T of edge devices and the relationship curve between uploading energy consumption E_u and transmission time T are shown in Figures 4 and 5, respectively. Under the three circumstances, both power and energy consumption decrease along with increasing restraint uploading time T. As proven by Lemma 2, a longer transmission time corresponds to a lower energy consumption. The proposed optimal strategy allocates additional bandwidth to those devices with a weak computing capability, thereby allowing them to complete the model parameter uploading and reduce the transmission power within the transmission restraint time. Figures 4 and 5 show that the proposed strategy reduces the transmission power by up to 21% and 23% compared with the average and random bandwidth allocations, respectively. Meanwhile, the transmission energy consumption increases by 1.4% and 1.9% compared with the baselines.

ii. Figures 6 and 7 compare the uploading power P_u and uploading energy consumption E_u in the three circumstances at each uplink gain h_{us}^j. At this time, the channel uplink gain h_{us}^j takes a random value within the range of $(10^{-5}, 2*10^{-5})$. Under the three circumstances, the transmission power and energy consumption decrease along with the improvement of channels. One observation from Corollary 1 is that more bandwidths should be allocated to those edge devices with weak channels. Overcoming such problem allows us to boost the transmission power or increase the bandwidths. The latter solution is preferred for energy minimization. As proven in Lemma 2, better channels correspond to lower transmission energy consumption. The transmission power in the proposed optimal strategy is 19% and 31% lower than those in the

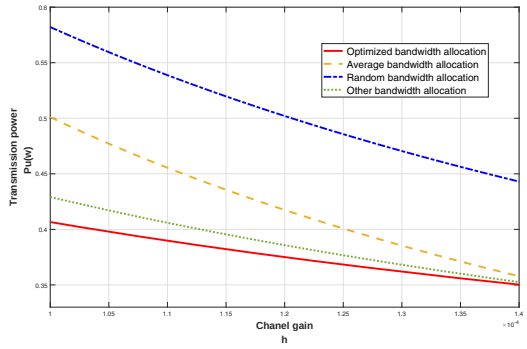


Fig. 6. Relationship between the uploading power P_u and uplink gain h_{us}^j of edge.

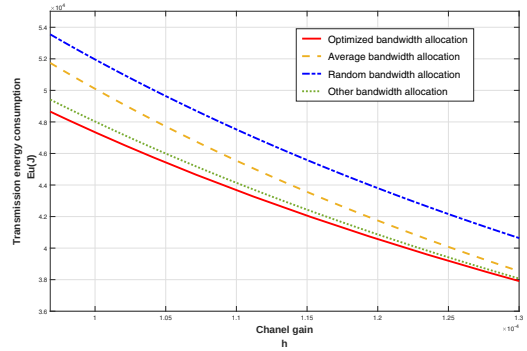


Fig. 7. Relationship between uploading energy consumption E_u and uplink gain h_{us}^j.

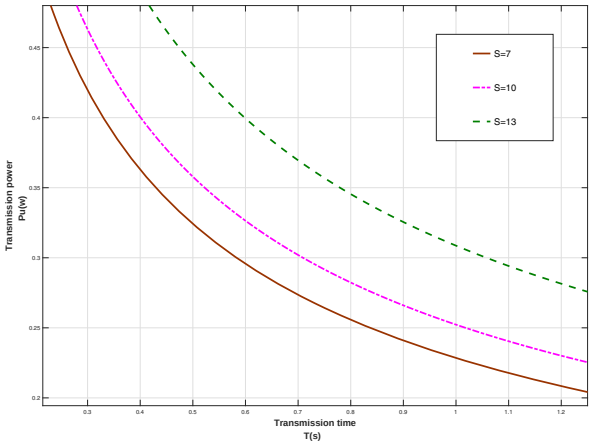


Fig. 8. Transmission time T and uploading power P_u under different multi - intelligent edge unit.

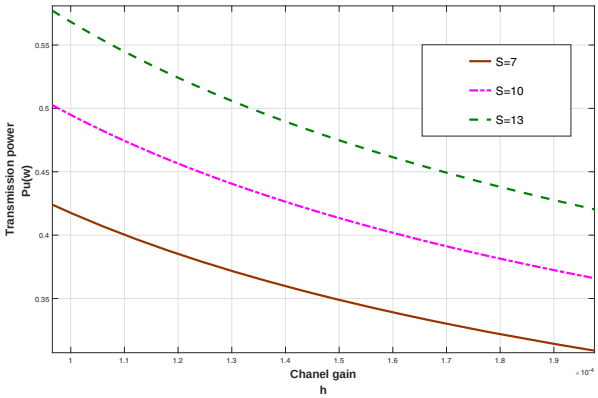


Fig. 9. Uplink gain h_{us}^j and uploading power P_u under different multi - intelligent edge unit.

average and random bandwidth allocations, respectively, and the improvements in transmission energy consumption are 5% and 9.6% higher than the baselines.

In order to further verify the effectiveness of the algorithm, in addition to the random and average bandwidth allocation strategies, we refer to the BW scheme in [27] as a control experiment. Under the channel state that we can tolerate, the maximum transmission power allocated by BW strategy and bandwidth optimization proposed in this paper are 0.43W and 0.41W respectively. The transmitted power of the proposed scheme will reduce 5.24%. In terms of transmission energy consumption, the proposed scheme is optimized by 1.54%.

iii. The purpose of reflecting the generality of the algorithm, we also carry out tests on intelligent edge units with varying numbers $(S = 10, S = 13)$. The addition of intelligent edge units will lead to more inter-cell interference during parameter transmission, and the terminals will increase transmission power to overcome this effect. As shown in figures 8 and 9, with the increase of the number of cells, the transmission power and energy consumption of equipment are significantly improved. More importantly, the experimental results show the universality of the proposed optimized bandwidth allocation. Even if the number of multiple edge cell is different, the proposed scheme can still effectively reduce the transmission power and energy consumption of the terminals according to the transmission delay limit and channel state.

VI. Conclusions

A low power consumption UPA strategy for FEL in a multi-unit intelligent edge network is proposed in this paper. To facilitate optimization, we divide UPA into the power allocation optimization strategy based on user QoE and the BA strategy. Solution seeking is performed by using the quasi-convex optimization and convex optimization techniques, and a low-complexity algorithm is proposed to solve the quasi-convex optimization problem. To satisfy user QoE, the proposed UPA strategy reduces the energy consumption of devices while adapting to the computing capability of channels and edge devices. Simulation results show that the proposed strategy outperforms the baselines in terms of transmission power and transmission energy consumption by 31% and 9.6%, respectively.

Appendix A

A. Proof of Lemma 1

First, we proof that $\Gamma_s(p_u)$ is quadratic differentiable on $\mathbb{R}$. Now we check the second-order condition of strictly quasi-convex function requiring p to satisfy $\Gamma_s^{'}(p) = 0$ and

$\Gamma_s^{''}(p) > 0$[40]. The first and second derivatives of $\Gamma_s(p_u)$ can be calculated as

$$\Gamma_s^{'}(p) = \frac{\psi_u C_u(p_u) - \frac{\vartheta_{us} D_u(p_u)}{A_u(p_u) ln2}}{C_u^2(p_u)}, \tag{20}$$

$$\Gamma_s^{''}(p) = \frac{\vartheta_{us}[G_{us}(p_u)C_{us}(p_u) + 2\vartheta_{us} D_{us}(p_u)/ln2]}{A_{us}^2(p_u) C_{us}^3(p_u) ln2}, \tag{21}$$

where,

$$A_{us}(p_u) = 1 + \vartheta_{us} p_u, \tag{21a}$$

$$C_{us}(p_u) = log_2(1 + \vartheta_{us} p_u), \tag{21b}$$

$$D_{us}(p_u) = \phi_u + \psi_u p_u, \tag{21c}$$

$$G_{us}(p_u) = \vartheta_{us} D_{us}(p_u) - 2\psi_u A_{us}(p_u). \tag{21d}$$

Suppose $\overline{p}_u \in (0, P_u]$, in order to prove $\Gamma_s^{'}(p) = 0$, we need,

$$\Omega_s(\overline{p}_u) = \psi_u log_2(1 + \vartheta_{us}\overline{p}_u) - \frac{\vartheta_{us}(\phi_u + \psi_u \overline{p}_u)}{(1 + \vartheta_{us}\overline{p}_u) ln2} = 0, \tag{22}$$

Substitute $\overline{p}_u$ into (21), we get,

$$\Gamma_s^{''}(\overline{p}_u) = \frac{\vartheta_{us}^3 D_{us}^2(\overline{p}_u)}{A_{us}^2(\overline{p}_u) C_{us}^3(\overline{p}_u) \psi_u ln^2 2}, \tag{23}$$

It can be easily verified that $\forall \overline{p}_u \in (0, P_u], \vartheta_{us}$ and $D_{us}^2(\overline{p}_u)$ are always positive. Therefore, $\Gamma_s^{''}(p) > 0$, and $\Gamma_s(p_u)$ are strictly quasiconvex functions on $(0, P_u]$.

B. Proof of Theorem 1

As mentioned above, $t_u^* = T_u, u \in U$.Next, we prove the optimal bandwidth allocation policy. Substitute $t_u = T_u$ into (18), it can be rewritten as

$$\min_{\delta_u, t_u} \sum_{u \in U} \frac{\delta_u B T_u(\tilde{I}_{us}^j + N_0)}{h_{us}^j} (2^{\frac{d_u}{\delta_u B T_u}} - 1), \tag{24}$$

$$s.t. \quad \sum_{u \in U} \delta_u = 1, 0 \le \delta_u \le 1 \quad . \tag{24a}$$

Since the above problem is a convex problem, by introducing Lagrange multipliers $\mu^* = [\mu_1^*, \mu_2^*, \cdots, \mu_U^*]^T \in \mathbb{R}^U$ for the inequality constraints $\delta \succeq 0$, with $\delta = [\delta_1, \delta_2, \cdots, \delta_U]^T$, and a multiplier $\upsilon^* \in \mathbb{R}$ for the equality constraint $1^T\delta = 1$, the KKT conditions can be written as follows

$$\delta^* \succeq 0, 1^T\delta^* = 1, \mu^* \succeq 0, \mu_u^*\delta_u^* = 0, u \in U,$$

$$\frac{B T_u(\tilde{I}_{us}^j + N_0)}{h_{us}^j} (2^{\frac{d_u}{\delta_u^* B T_u}} - \frac{d_u ln2}{\delta_u^* B T_u} 2^{\frac{d_u}{\delta_u^* B T_u}} - 1) - \mu_u^* + \upsilon^* = 0, u \in U. \tag{25}$$

By solving the above equations, we can get

$$\delta_u^* = \frac{d_u ln2}{B T_u[1 + \mathcal{W}(\frac{h_{us}^j \upsilon^* - B T_u(\tilde{I}_{us}^j + N_0)}{B T_u(\tilde{I}_{us}^j + N_0)})]} \tag{26}$$

where $\mathcal{W}(\cdot)$ is the Lambert W function, and the Lagrange multiplier value υ^* is obtained by solving $\sum_{k=1}^K [1 + \mathcal{W}(\frac{h_{us}^j \upsilon^* - B T_u(\tilde{I}_{us}^j + N_0)}{B T_u(\tilde{I}_{us}^j + N_0)})] = 1$.

C. Proof of Corollary 1

First, we prove that δ_u^* is a non-increasing function about T_u.Denote $x = \frac{h_{us}^j \upsilon^* - B T u(\tilde{I}_{us}^j + N_0)}{B T u(\tilde{I}_{us}^j + N_0) e}$,then we can obtain $T_u = \frac{h_{us}^j \upsilon^*}{(x + \frac{1}{e}) B(\tilde{I}_{us}^j + N_0) e}$.Substituting it to the expression for γ_k^*, one can we have,

$$\begin{aligned} \delta_u^* &= \frac{d_u ln2}{B T_u[1 + \mathcal{W}(\frac{h_{us}^j \upsilon^* - B T_u(\tilde{I}_{us}^j + N_0)}{B T_u(\tilde{I}_{us}^j + N_0)})]} \\ &= \frac{(\tilde{I}_{us}^j + N_0) e d_u ln2}{h_{us}^j \upsilon^*} \frac{x + \frac{1}{e}}{1 + \mathcal{W}(x)}. \end{aligned} \tag{39}$$

Further, we denote,

$$y = \frac{x + \frac{1}{e}}{1 + \mathcal{W}(x)} = \frac{\mathcal{W}(x) e^{\mathcal{W}(x)} + \frac{1}{e}}{1 + \mathcal{W}(x)}. \tag{40}$$

It can be easily proved that y is a non-decreasing function with respect to $\mathcal{W}(x)$. Since $\mathcal{W}(x)$ is a non-decreasing function of x, $x(T_k)$ is a non-increasing function of T_k, it follows that δ_u^* is non-increasing of T_u.

Next, we prove that δ_u^* is a non-increasing function with respect to h_{us}^j.$h_{us}^j = \frac{B T_u(\tilde{I}_{us}^j + N_0) e}{\upsilon^*}(x + \frac{1}{e})$can be obtained from $x = \frac{h_{us}^j \upsilon^* - B T_u(\tilde{I}_{us}^j + N_0)}{B T_u(\tilde{I}_{us}^j + N_0) e}$, Substituting it into the expression for δ_u^*, it follows that

$$\delta_u^* = \frac{d_u ln2}{B T_u[1 + \mathcal{W}(\frac{h_{us}^j \upsilon^* - B T_u(\tilde{I}_{us}^j + N_0)}{B T_u(\tilde{I}_{us}^j + N_0)})]} = \frac{d_u ln2}{B T_u} \frac{1}{1 + \mathcal{W}(x)}. \tag{41}$$

Further, we let

$$z = \frac{1}{1 + \mathcal{W}(x)}. \tag{42}$$

Obviously, z is non-increasing with respect to $\mathcal{W}(x)$. Because $\mathcal{W}(x)$ is non- decreasing about x, $x(h_{us}^j)$ is non-decreasing about h_{us}^j, we can conclude that δ_u^* is a non-increasing function with respect to h_{us}^j.

REFERENCES

[1] M. Chen, U. Challita, W. Saad, C. Yin, and M. Debbah, "Artificial neural networks-based machine learning for wireless networks: A tutorial," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3039–3071, Jul. 2019.

[2] Y. Sun, M. Peng, Y. Zhou, Y. Huang, and S. Mao, "Application of machine learning in wireless networks: Key techniques and open issues," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3072–3108, Jun. 2019.

[3] Y. Liu, S. Bi, Z. Shi, and L. Hanzo, "When machine learning meets big data: A wireless communication perspective," *arXiv preprint arXiv:1901.08329*, Jan. 2019.

[4] E. Li, L. Zeng, Z. Zhou, and X. Chen, "Edge AI: On-demand accelerating deep neural network inference via edge computing," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp.447–457, Oct. 2019.

[5] X. Wang, Y. Han, C. Wang, Q. Zhao, X. Chen, and M. Chen, "In-Edge AI: Intelligentizing mobile edge computing, caching and communication by federated Learning," *IEEE Netw.*, vol. 33, no. 5, pp. 156–165, Jul. 2019.

[6] G. Zhu, D. Liu, Y. Du, C. You, J. Zhang, and K. Huang, "Towards an intelligent edge: Wireless communication meets machine learning," *IEEE Commun. Mag.*, vol. 58, no. 1, pp. 19–25, Jan. 2019.

[7] D. Wen, X. Li, Q. Zeng, J. Ren, and K. Huang, "An overview of data-importance aware radio resource management foredge machine learning," *J. Commun. Inf. Netw.*, vol. 4, no. 4, pp. 1–14, Dec. 2019.

[8] K. Bonawitz *et al.*, "Towards federated learning at scale: System design," *arXiv preprint arXiv:1902.01046*, Mar. 2019.

[9] L. WANG, W. WANG and B. LI, "CMFL: Mitigating Communication Overhead for Federated Learning," in *Proc. IEEE ICDCS*, 2019.

[10] S. Ha, J. Zhang, O. Simeone, and J. Kang, "Coded federated computing in wireless networks with straggling devices and imperfect CSI," *arXiv preprint arXiv:1901.05239*, Jan. 2019.

[11] O. Habachi, M. A. Adjif, and J. P. Cances, "Fast uplink grant for NOMA: A federated learning based approach," *arXiv preprint arXiv:1904.07975*, Mar. 2019.

[12] J. Chen, X. Pan, R. Monga, S. Bengio, and R. Jozefowicz, "Revisiting distributed synchronous sgd," *arXiv preprint arXiv:1604.00981*, [Online]. Available: https://arxiv.org/abs/1604.00981, 2017.

[13] Cui L, X Su, Ming Z, et al, "Blockchain-assisted compression algorithm of federated learning for content caching in edge computing," *IEEE Internet Things J.*, Early Access, Aug. 2020.

[14] Tra Huong Thi Le, Nguyen H. Tran, Yan Kyaw Tun, Zhu Han and Choong Seon Hong, "Auction based incentive design for efficient federated learning in cellular wireless networks," in *Proc. IEEE WCNC*, 2020.

[15] Laizhong Cui, Xiaoxin Su, Yipeng Zhou, Yi Pan. "Slashing communication traffic in federated learning by transmitting clustered model updates," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 8, pp.2572–2589, Aug. 2021.

[16] Nguyen H. Tran, Wei Bao, Albert Zomaya, Minh N. H. Nguyen and Choong Seon Hong, "Federated learning over wireless networks: Optimization model design and analysis," in *Proc. IEEE INFOCOM*, 2019.

[17] M. Chen, Z. Yang, W. Saad, C. Yin, H. V. Poor and S. Cui, "A joint learning and communications framework for federated learning over wireless networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 1, pp. 269–283, Jan. 2021.

[18] X. Lyu, H. Tian, P. Zhang, and C. Sengul, "Multi-user joint task offloading and resources optimization in proximate clouds," *IEEE Trans. Veh. Technol.*, vol. 66, no. 4, pp. 3435–3447, Apr. 2017.

[19] L.Yang, J.Cao, H.Cheng, andY.Ji, "Multi-user computation partitioning for latency sensitive mobile cloud applications," *IEEE Trans. Comput.*, vol. 64, no. 8, pp. 2253–2266, Aug. 2015.

[20] S. Bi, J. Lyu, Z. Ding, and R. Zhang, "Engineering radio maps for wireless resource management," *IEEE Wireless Commun.*, vol. 26, no. 2, pp. 133–141, Apr. 2019.

[21] W. Liu, J. Wei and Q. Meng, "Comparisions on KNN, SVM, BP and the CNN for handwritten digit recognition," in *Proc. IEEE AEECA*, 2020, pp. 587-590.

[22] X. Mei, Q. Wang, and X. Chu, "A survey and measurement study of GPU DVFS on energy conservation," *Digital Commun. Netw.*, vol. 3, no. 2, pp. 89–100, 2017.

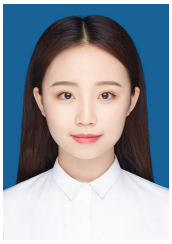
[23] E. Dahlman, S. Parkvall, and J. Skold, "4G:LTE/LTE-Advanced for Mobile Broadband," New York, NY, USA: Academic, 2013.

[24] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "User association for load balancingin heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2706–2716, Jun. 2013.

[25] S. Boyd and L. Vandenberghe, *Convex Optimization,* Cambridge, U.K.: Cambridge Univ. Press, 2004.

[26] B. Bereanu, "Quasi-convexity, strictly quasi-convexity and pseudoconvexity of composite objective functions," *Revue Franc¸aise D'automatique, Informatique, Recherche Operationnelle Mathematique*, vol. 6, no. 1, pp. 15–26, 1972.

[27] Chen J , H Xing, Lin X , et al. "Joint Cache Placement and Bandwidth Allocation for FDMA-based Mobile Edge Computing Systems," in *Proc. IEEE ICC*, 2020.



Tianyi Zhou received the B.E. degree in Information and Communication Engineering from from Beijing Information Science and Technology University, China, in 2019. She is currently pursuing the M.Phil. degree with the School of Information and Communication Engineering, Beijing Information Science and Technology University. Her research interests include wireless communication (5G), federated learning and intelligent resource management.

Xuehua Li received the Ph.D. degree in telecommunications engineering from the Beijing University of Posts and Telecommunications, Beijing, China, in 2008. She is currently a Professor and the Deputy Dean of the School of Information and Communication Engineering with Beijing Information Science and Technology University, Beijing. She is a Senior Member of the Beijing Internet of Things Institute. Her research interests are in the broad areas of communications and information theory, particularly the Internet of Things, and coding for multimedia communications system.

Chunyu Pan received the Ph.D. degree with the School of Information and Communication Engineering, Beijing University of Posts and Telecommunications. Since 2019, she has been in Beijing Information Science and Technology University, where she is an Associate Professor in school of Information and Communication Engineering. Her main research interests include mobile communications and future networks, intelligent resource management, and UAV communications.

Mingyu Zhou received the Ph.D. degree in Information and Communication Engineering from Beijing University of Posts and Telecommunications.He focus on innovation related to future wireless communication technologies. He has working experience with more than 10 years. He has released more than 20 papers and applied more than 100 patents.

Yuanyuan Yao received the Ph.D. degree in Information and Communication Engineering from Beijing University of Posts and Telecommunications, Beijing, China, in 2017. Since 2017, she has been with the School of Information and Communication Engineering, Beijing Information Science and Technology University, Beijing, China, as an Associate Professor. Her research interests include UAV communications, stochastic geometry and its applications in large-scale wireless networks, energy harvesting.