Joint User Selection and Resource Allocation for Fast Federated Edge Learning

JIANG Zhihui, HE Yinghui, YU Guanding

(College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, Zhejiang 310027, China)

Abstract: By periodically aggregating local learning updates from edge users, federated edge learning (FEEL) is envisioned as a promising means to reap the benefit of local rich data and protect users' privacy. However, the scarce wireless communication resource greatly limits the number of participated users and is regarded as the main bottleneck which hinders the development of FEEL. To tackle this issue, we propose a user selection policy based on data importance for FEEL system. In order to quantify the data importance of each user, we first analyze the relationship between the loss decay and the squared norm of gradient. Then, we formulate a combinatorial optimization problem to maximize the learning efficiency by jointly considering user selection and communication resource allocation. By problem transformation and relaxation, the optimal user selection policy and resource allocation are derived, and a polynomial-time optimal algorithm is developed. Finally, we deploy two commonly used deep neural network (DNN) models for simulation. The results validate that our proposed algorithm has strong generalization ability and can attain higher learning efficiency compared with other traditional algorithms.

Keywords: data importance; federated edge learning; learning accuracy; learning efficiency; resource allocation; user selection

DOI: 10.12142/ZTECOM.202002004

https://kns.cnki.net/kcms/detail/34.1294.TN.20200612.1311.003.html, published online June 12, 2020

Manuscript received: 2020-01-31

Citation (IEEE Format): Z. H. Jiang, Y. H. He, and G. D. Yu, "Joint user selection and resource allocation for fast federated edge learning," *ZTE Communications*, vol. 18, no. 2, pp. 20–30, Jun. 2020. doi: 10.12142/ZTECOM.202002004.

1 Introduction

With the explosive growth of data generated by mobile devices and the remarkable breakthroughs made in artificial intelligence (AI) in recent years, the combination of AI and wireless networks is attracting more and more interests[1]. To leverage the abundant data, which are unevenly distributed over a large number of edge devices, and to train a high quality prediction model, the traditional scheme is to do centralized learning by transmitting the raw data to the data center. However, this scheme has two drawbacks. On the one hand, the privacy of users may be divulged when the data center suffers from malicious attacks. On the other hand, the communication latency is long since the volume of data is large and the communication resource is limited. To overcome these two issues, a new framework, namely federated edge learning (FEEL), has been recently proposed in Ref. [2]. This framework makes a collaboration of the distributed learning framework, named federated learning (FL)

This work was supported in part by the National Natural Science Foundation of China under Grant No. 61671407.

[3] and mobile edge computing (MEC)[4], which not only ensures users' privacy but also exploits the computing resource of both edge devices and edge servers.

In the FEEL system, edge devices need to interact with the edge server constantly to train a global model. Thus, communication cost is one of the major constraints of model training since the wireless communication resource is limited. Recently, several works have investigated accelerating the training task by reducing the communication overhead[5–6]. To achieve a low-latency FEEL system, the authors in Ref. [5] propose a broadband analog aggregation scheme by exploiting over-the-air computation and derive two communication-and-learning tradeoffs. In Ref. [6], the authors propose a new protocol to reduce the communication overhead and improve the training speed by selecting devices as many as possible based on their channel state information (CSI). Besides, energy-efficient FL over wireless networks has been investigated in Refs. [7] and [8]. In Ref. [7], energy-efficient strategies are proposed for joint bandwidth allocation and energy-and-learning aware scheduling with less energy consumption. The authors in Ref. [8] propose an iterative algorithm to achieve the tradeoff between latency and energy consumption for FL. Moreover, several recent works focus on the problem of user selection for FL over wireless networks[9–12]. In Ref. [9], the authors derive a tradeoff between the number of scheduled users and subchannel bandwidth under fixed amount of available spectrum. To improve the running efficiency of FL, the authors in Ref. [10] propose a scheduling policy by exploiting the CSI, i.e., the instantaneous channel qualities. In Ref. [11], the authors consider a user selection problem based on packet errors and the availability of wireless resources, and a probabilistic user selection scheme is proposed to reduce the convergence time of FL in Ref. [12].

However, the aforementioned works ignore the fact that the process of model training is time-consuming as well. According to Ref. [13], different training samples are not equally important in a training task. Therefore, faced with the massive data, the topic of selecting important data to further accelerate the training task deserves to be studied. Several recent works have studied on this topic. In Refs. [14] and [15], data importance is quantified by the signal-to-noise ratio (SNR) and data uncertainty measured by the distance to the decision boundary. Based on this, the authors propose a data importance aware retransmission protocol and a user scheduling algorithm, respectively.

As we have mentioned before, some works have already investigated the acceleration of the training task based on data importance. However, this topic has not been investigated in the FEEL system yet, which is a distributed edge learning system. Inspired by this, we consider an FEEL system, where the learning efficiency of the system is improved by user selection based on data importance. First, we analyze the relation between the loss decay and the learning update information (LUI), i.e., the squared norm of the gradient, and derive an indicator to quantify the data importance. Then, an optimization problem to maximize the learning efficiency of the FEEL system is formulated by joint user selection and communication resource allocation. The closed-form solution for optimal user selection policy and communication resource allocation is derived by problem transformation and relaxation. Based on this, we develop a polynomial-time algorithm to solve this mixed-integer programming problem. Finally, we verify the generalization ability and the performance improvement of our proposed algorithm by extensive simulation.

The rest of this paper is organized as follows. In Section 2, we introduce the FEEL system and establish the deep neural network (DNN) model and communication model. In Section 3, we propose an indicator to quantify the data importance, analyze the end-to-end latency in each communication round, and formulate the optimization problem to maximize the learning efficiency. The optimal solution and the optimal algorithm are developed in Section 4. Simulation results are presented in Section 5 and the whole paper is concluded in Section 6.

2 System Model

In this section, we will first introduce the FEEL system model. Then, both the DNN model and communication model are introduced.

2.1 Federated Edge Learning System

We consider an FEEL system as shown in **Fig. 1**, which comprises an edge server and K distributed users, denoted by $\mathcal{K} = \{1,2,...,K\}$. Each user utilizes its local dataset to train the local DNN model. Let $\mathcal{D}_k = \{(\mathbf{x}_1,y_1),...,(\mathbf{x}_{N_k},y_{N_k})\}$ denote the local dataset of user k, where $\mathbf{x}_i$ is the training sample, y_i is the size of the corresponding ground-true label, and N_k is the size of dataset. During each communication round, users first upload their gradients to the edge server. Then, the edge server collects the local gradients from users and aggregates them as the global gradient. Users update their local models by the global gradient broadcast by the edge server. Ultimately, users are supposed to collaborate with each other in training a shared global model. Therefore, users' privacy is protected since the raw data are not transmitted to the edge server. However, due to the limited wireless communication resource, the number of users participated in the training task is restricted. To tackle this issue, we intend to propose a user selection policy by jointly considering the LUI and CSI of each user. During each communication round, users' data are not of equal importance. So we only select part of users to upload their local gradients based on data importance and channel data rate. The following seven steps are defined as a communication round.

1) Calculate local gradient. In the n-th communication round, each user utilizes its local dataset to train its local mod-

el. Denote $\boldsymbol{\theta}$ as the parameter set of the DNN model. The local gradient vector $\boldsymbol{G}_k^{\boldsymbol{\theta}}[n]$ can be calculated by the backpropagation algorithm. Note that the local model and the local gradient are different among users since different users may have different datasets.

2) Upload the squared norm of local gradient. After obtaining the local gradient vector $\boldsymbol{G}_k^{\boldsymbol{\theta}}[n]$, each user calculates the squared norm of local gradient $\left\|\boldsymbol{G}_k^{\boldsymbol{\theta}}[n]\right\|_2^2$ and transmits it to the edge server. Here, $\|\cdot\|_2$ is the L2 norm.

3) Select User. The edge server receives the squared norm of local gradients from all users. Based on data importance and channel data rate, the edge server will determine which users are going to be selected to participate in the training task.

4) Upload local gradient of selected users. In this step, those selected users upload their local gradients to the edge server via the time division multiple access (TDMA) method without loss of generality.

5) Aggregate global gradient. The edge server receives the local gradients of all selected users and then aggregates them as the global gradient, which can be expressed as

$$\boldsymbol{G}^{\boldsymbol{\theta}}[n] = \frac{1}{\left|\bigcup_k a_k \mathcal{D}_k\right|} \sum_{k=1}^{K} a_k \left|\mathcal{D}_k\right| \boldsymbol{G}_k^{\boldsymbol{\theta}}[n], \tag{1}$$

where $a_k \in \{0, 1\}$ indicates whether user k is selected, i.e., $a_k = 1$ if user k is selected and $a_k = 0$ otherwise.

6) Broadcast global gradient. After finishing the global gradient aggregation, the base station (BS) broadcasts the global gradient to all the users.

7) Update local model. After the global gradient is received, each user updates its local model, as

$$\boldsymbol{\theta}[n+1] = \boldsymbol{\theta}[n] - \xi[n]\boldsymbol{G}^{\boldsymbol{\theta}}[n], \tag{2}$$

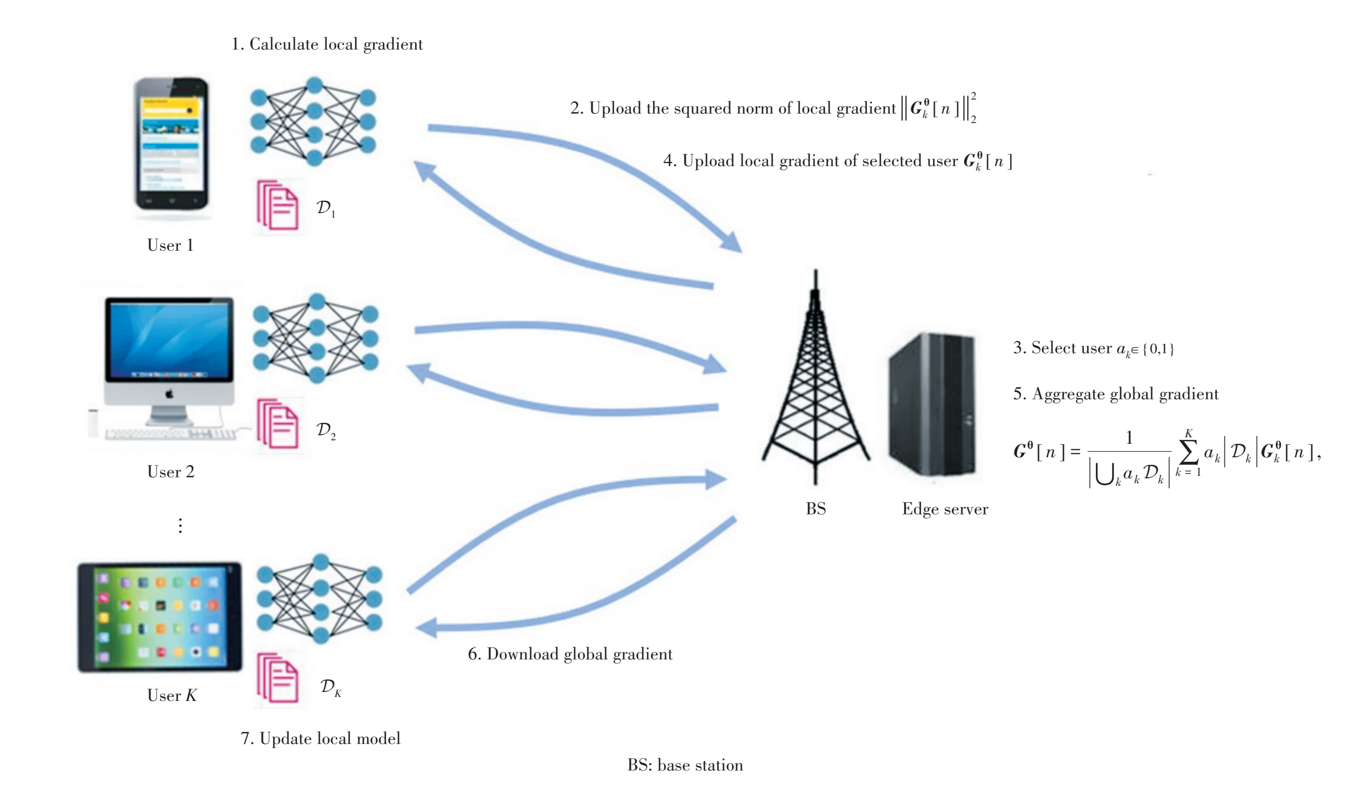
where $\xi[n]$ is the learning rate of the n-th communication round.

The above seven steps are periodically performed until the global model converges. During the training process, the local gradient and the CSI of users are different in each communication round. Therefore, the edge server should run the optimal algorithm to select users in each communication round.

2.2 DNN Model

In this work, all users adopt the same DNN model for training. To evaluate the error between the learning output and the ground-true label y_i, we define the loss function of training samples as $l(\boldsymbol{\theta},\mathbf{x}_i,y_i)$. Thus, the local loss function of user k and the global loss function can be represented as

$$L_k(\boldsymbol{\theta},\mathcal{D}_k) = \frac{1}{\left|\mathcal{D}_k\right|} \sum_{(\mathbf{x}_i,y_i)\in\mathcal{D}_k} l(\boldsymbol{\theta},\mathbf{x}_i,y_i), \tag{3}$$


1. Calculate local gradient
User 1
$\mathcal{D}_1$
User 2
$\mathcal{D}_2$
⋮
User K
$\mathcal{D}_K$
7. Update local model
2. Upload the squared norm of local gradient $\left\|\boldsymbol{G}_k^{\boldsymbol{\theta}}[n]\right\|_2^2$
4. Upload local gradient of selected user $\boldsymbol{G}_k^{\boldsymbol{\theta}}[n]$
6. Download global gradient
BS
Edge server
3. Select user $a_k \in \{0,1\}$
5. Aggregate global gradient
$\boldsymbol{G}^{\boldsymbol{\theta}}[n] = \frac{1}{\left|\bigcup_k a_k \mathcal{D}_k\right|} \sum_{k=1}^{K} a_k \left|\mathcal{D}_k\right| \boldsymbol{G}_k^{\boldsymbol{\theta}}[n],$
BS: base station

▲**Figure 1. Seven steps in each communication round.**

$$L(\boldsymbol{\theta}) = \frac{1}{\left|\bigcup_k a_k \mathcal{D}_k\right|} \sum_{k=1}^{K} a_k \left|\mathcal{D}_k\right| L_k(\boldsymbol{\theta}, \mathcal{D}_k), \tag{4}$$

respectively. In the course of training, the global loss function $L(\boldsymbol{\theta})$ is the objective function to be minimized. In our scheme, we aim to accelerate the training task and train a high quality global model. Without loss of generality, we utilize stochastic gradient descent (SGD) as the optimal algorithm. Then, the local gradient vector of user k is given by

$$\boldsymbol{G}_k^{\boldsymbol{\theta}} = \nabla L_k(\boldsymbol{\theta}, \mathcal{D}_k), \tag{5}$$

where ∇ implies the gradient operator.

2.3 Communication Model

As described above, distributed users and the edge server need to exchange data from each other in each communication round. In our scheme, two frequently-used approaches of data transmission are adopted, named TDMA and broadcasting.

First, those selected users upload their local gradients to the edge server via the TDMA method. Specifically, a time frame is divided into n time slots. Each user transmits its data on its own time slot. According to Ref. [16], the length of each time frame in LTE standards is 10 ms. Actually, the transmission delay of the gradients is on the scale of second, which is far larger than the length of a time frame[17]. Therefore, we can use the average uplink channel capacity, rather than the instantaneous channel capacity, to evaluate the data rate of user k[18], which can be expressed as

$$R_k^U = W \mathbb{E}_h \left\{ \log_2 \left(1 + \frac{p_k^U \left| h_k^U \right|^2}{N_0} \right) \right\}, \tag{6}$$

where h_k^U is the uplink channel power gain of user k, p_k^U is the corresponding transmission power, $\mathbb{E}_h$ is the expectation over the uplink channel power gain, W is the system bandwidth, and N_0 is the noise power.

After the global gradient aggregation is finished, the BS will broadcast the global gradient to all users. In this way, all users are able to receive the global gradient synchronously. Let h_k^D denote the downlink channel power gain of user k and p^D denote the transmission power for all users. Thus, the downlink data rate is given by

$$R^D = W \min_{k \in \mathcal{K}} \left\{ \mathbb{E}_h \left\{ \log_2 \left(1 + \frac{p^D \left| h_k^D \right|^2}{N_0} \right) \right\} \right\}. \tag{7}$$

3 Problem Formulation

In this section, we will first propose an indicator to quantify the data importance of users. Then, we analyze the end-to-end latency in each communication round and formulate the optimization problem to maximize the lower bound of the system learning efficiency.

3.1 Importance Analysis

In each communication round, only part of users is selected to participate in the training task because of the limited wireless communication resource. According to Ref. [13], different training samples do not equally contribute to the model training. Consequently, we intend to select users based on the level of data importance as well as the channel data rate. To quantify the data importance, we define the loss decay function as

$$\Delta L[n] = L(\boldsymbol{\theta}[n-1]) - L(\boldsymbol{\theta}[n]). \tag{8}$$

The loss decay function $\Delta L[n]$ indicates the decrease of the loss in the n-th communication round. From Eq. (8), in the same period of time, the larger the loss decays, the faster the training speed is. In other words, the loss decay reflects the data importance to some extent.

According to Ref. [19], the loss decay is proportional to the squared norm of the gradient. Thus, the lower bound of the loss decay in the n-th communication round is given as

$$\Delta L[n] \geqslant \beta \left\| \boldsymbol{G}^{\boldsymbol{\theta}}[n] \right\|_2^2, \tag{9}$$

where β is a constant determined by the learning rate and the specific DNN model. Therefore, we can further link the data importance with the squared norm of the gradient vector. With the above discussions, we can quantify the data importance of user k by the squared norm of its local gradient, which can be represented as

$$\rho_k = \beta \left\| \boldsymbol{G}_k^{\boldsymbol{\theta}}[n] \right\|_2^2, \forall k \in \mathcal{K}. \tag{10}$$

Therefore, the lower bound of the global loss decay in a communication round can be expressed as

$$\Delta L = \sum_{k=1}^{K} a_k \rho_k. \tag{11}$$

3.2 End-to-End Latency Analysis

As mentioned before, our goal is to improve the learning efficiency of the FEEL system. Thus, the end-to-end latency of one communication round should be optimized. The detailed analysis of latency in one communication round is given as follows.

1) Calculate local gradient. The latency of local training for user k is denoted by T_k^L.

2) Upload local gradient of selected users. As we mentioned before, only those selected users upload their local gradients to the edge server via TDMA. So the average transmission delay of user k can be expressed by

$$T_k^T = a_k \frac{V}{\tau_k R_k^U}, \forall k \in \mathcal{K}, \tag{12}$$

where τ_k is the proportion of the time slot for user k in a time frame and V is the volume of the gradient, which is a constant for all users.

3) Broadcast global gradient. For all users, the latency of downloading the global gradient is given by

$$T^D = \frac{V}{R^D}. \tag{13}$$

4) Update local model. Let us denote T_k^U as the delay of model updating for user k.

Since the squared norm of local gradient and the value of a_k are small enough, the corresponding transmission delay can be neglected. Besides, the edge server has powerful computing capacity in general. Therefore, the aggregation delay can also be neglected.

Then we provide further analysis to obtain the whole latency of one communication round. Note that all users receive the global gradient and start to update the local model synchronously. However, the delay of model updating and training varies since users may have different computing power. Hence, users are only allowed to upload the squared norm of local gradient to the edge server until they all finish model updating and training. In addition, the edge server should begin to aggregate the global gradient until those selected users have uploaded their local gradients. Based on the above analysis, the end-to-end latency of the FEEL system in one communication round is given by

$$T = \max_{k \in \mathcal{K}} \{T_k^U + T_k^L\} + \max_{k \in \mathcal{K}} T_k^T + T^D. \tag{14}$$

3.3 Problem Formulation

In this work, we aim to improve the learning efficiency of the FEEL system by jointly considering user selection and communication resource allocation. According to Ref. [20], we adopt the following criterion to evaluate the training performance of the FEEL system.

Definition 1: The learning efficiency of the FEEL system can be defined as

$$E = \frac{\Delta L}{T}. \tag{15}$$

Remark 1: The definition of the learning efficiency implies the decay rate of the global loss in a given time period T. The improvement of the learning efficiency means the acceleration of the training task. Therefore, it is appropriate to evaluate the training performance of the FEEL system by the learning efficiency. In our work, we aim to reduce the communication delay of each communication round. Besides, we maximize the lower bound of the system learning efficiency. Consequently, the learning efficiency of the FEEL system can be improved.

Based on the above analysis, the optimization problem can be mathematically formulated as

$$\mathcal{P}1: \max_{\{a_k, \tau_k, T\}} E = \frac{\Delta L}{T} = \frac{\sum_{k=1}^{K} a_k \rho_k}{T}, \tag{16a}$$

$$s.t. \max_{k \in \mathcal{K}} \{T_k^U + T_k^L\} + T_k^T + T^D \leqslant T, \forall k \in \mathcal{K}, \tag{16b}$$

$$\sum_{k=1}^{K} \tau_k \leqslant 1, \tag{16c}$$

$$a_k \in \{0,1\}, \forall k \in \mathcal{K}, \tag{16d}$$

$$\tau_k, T \geqslant 0, \forall k \in \mathcal{K}, \tag{16e}$$

where the constraint (16b) indicates that the end-to-end latency of each user in one communication round is no more than the end-to-end latency of the FEEL system and the constraint (16c) represents the uplink communication resource limitation. For description convenience, we rewrite $\max_{k \in \mathcal{K}} \{T_k^U + T_k^L\} + T^D$ as T^C in the following sections.

4 Optimal Solution

4.1 Problem Transformation

It is evident that the optimization problem $\mathcal{P}1$ is a mixed-integer programming problem. Since the objective function of $\mathcal{P}1$ is non-convex, it is rather challenging to directly solve it. Combining Eqs. (12) and (16b), we notice that T is relevant to a_k and τ_k. When a_k and τ_k are fixed, the variable T must be minimized to maximize the learning efficiency. Therefore, the optimal solution to problem $\mathcal{P}1$ can be obtained when "$\leqslant$" in the constraint (16b) is set to "=", i.e. $\tau_k = a_k V / R_k^U (T - T^C)$.

However, problem $\mathcal{P}1$ is still hard to solve due to the integer constraint (16d). Therefore, we relax the integer constraint $a_k \in \{0,1\}$ to the real-value constraint $a_k \in [0,1]$. Problem $\mathcal{P}1$ can then be relaxed into problem $\mathcal{P}2$, which is given by

$$\mathcal{P}2: \max_{\{a_k, T\}} \frac{\sum_{k=1}^{K} a_k \rho_k}{T}, \tag{17a}$$

$$s.t. \sum_{k=1}^{K} \frac{a_k V}{R_k^U} \leqslant T - T^C, \tag{17b}$$

$$a_k \in [0,1], \forall k \in \mathcal{K}, \tag{17c}$$

$$T \geqslant 0. \tag{17d}$$

In the following sections, we first obtain the optimal solution to problem $\mathcal{P}2$ with fixed T. Then, we continue to solve the problem $\mathcal{P}2$ with varying T, and the optimal solution to problem $\mathcal{P}1$ is finally derived.

4.2 Optimal User Selection

We now solve the problem $\mathcal{P}2$. When T is given, problem $\mathcal{P}2$ can be converted to a standard convex optimization problem since the objective function is concave and all constraints are convex. Thus, we can derive the optimal solution to $\mathcal{P}2$ with fixed T.

Theorem 1: The optimal solution to problem $\mathcal{P}2$ with fixed T is given as follows.

1) If $\rho_k R_k^U < \lambda^*, a_k^* = 0$;

2) If $\rho_k R_k^U > \lambda^*, a_k^* = 1$;

3) If $\rho_k R_k^U = \lambda^*, 0 \leqslant a_k^* \leqslant 1$,

where λ^* is the optimal value of the Lagrange multiplier satisfying the constraint (17b). Particularly, the real-value of a_k^* depends on the constraint (17b) if $\rho_k R_k^U = \lambda^*$.

Proof: See Appendix A.

Remark 2 (Optimal user selection policy): According to Theorem 1, λ^* can be regarded as the threshold which determines whether to select the user. Besides, the selection priority of user k depends on the product of its data importance ρ_k and the uplink data rate R_k^U. On the one hand, a user with more important data contributes more to the global model training. On the other hand, the transmission delay can be shortened by selecting users with higher uplink data rates. Thus, the system prefers to select users with larger values of $\rho_k R_k^U$. By doing so, the learning efficiency of the FEEL system can be improved.

4.3 Optimal System Latency and Communication Resource Allocation

In this part, we proceed to obtain the optimal system latency and develop the optimal communication resource allocation to further improve the learning efficiency of the FEEL system. So far, we have obtained the optimal user selection strategy when the system latency is invariant. Based on this, the optimal system latency must be obtained when "⩽" in the constraint (17b) is set to "=", i.e., $T = \sum_{k=1}^{K} a_k V / R_k^U + T^C$. In order to develop the optimal T and τ_k, we introduce the following theorem.

Theorem 2: The optimal solutions to problem $\mathcal{P}2$ and problem $\mathcal{P}1$ are exactly the same.

Proof: See Appendix B.

Remark 3 (Optimal system latency and communication resource allocation): Theorem 2 indicates that the optimal solution of a_k to problem $\mathcal{P}2$ must be an integer solution. Based on this, the range of feasible solutions to problem $\mathcal{P}2$ can be reduced greatly. Thus, we only need to compare the learning efficiency of the FEEL system when the total number of selected users varies. Here, users in the system are selected by the optimal user selection policy as aforementioned. So T^* that achieves the maximum learning efficiency is the optimal system latency to both problems $\mathcal{P}2$ and $\mathcal{P}1$, which can be expressed as

$$T^* = \sum_{k=1}^{K} \frac{a_k^* V}{R_k^U} + T^C. \tag{18}$$

As we have indicated before, when "⩽" in the constraint (16b) is set to "=", the solution must be the optimal solution of problem $\mathcal{P}1$. Consequently, we can obtain the optimal communication resource allocation by simple mathematical calculation, as

$$\tau_k^* = a_k^* \frac{V}{R_k^U (T^* - T^C)}. \tag{19}$$

The result in Eq. (19) shows that a less time slot is allocated for the user with a higher uplink data rate.

4.4 Optimal Algorithm for Problem $\mathcal{P}1$

Thus far, we have obtained the optimal solution to problem $\mathcal{P}1$. In this part, we intend to develop an optimal algorithm for problem $\mathcal{P}1$ based on the above analysis. As mentioned before, in order to obtain the optimal solution to problem $\mathcal{P}1$, all selection cases should be compared. However, this would become very time-consuming as the number of users increases. Therefore, a low computational complexity algorithm is required. We define E_M, $M \in \{1, 2, ..., K\}$ as the learning efficiency of the FEEL system when M users are selected. To better fit the practical systems, we have the following theorem.

Theorem 3: E_M increases first and then decreases with the increase of M.

Proof: See Appendix C.

Remark 4: Theorem 3 indicates that the learning efficiency E_M has only one global optimal. Therefore, we can select users successively by the optimal user selection policy until the learning efficiency of the FEEL system begins to decrease. By doing so, we are able to find the optimal solution to problem $\mathcal{P}1$. According to the above analysis, the optimal algorithm for problem $\mathcal{P}1$ is shown in **Algorithm 1**. We can easily find that the computational complexity of this algorithm is determined by the sort operation. Therefore, the computational complexity is $\mathcal{O}(K \log K)$. With regard to mixed-integer programming problems, it is acceptable to find the optimal solution with a polynomial-time complexity, indicating that this algorithm can be applied to practical systems.

Algorithm 1: The optimal algorithm for problem $\mathcal{P}1$

```
Calculate $\rho_k R_k^U, \forall k \in \mathcal{K}$.
Sort $\rho_k R_k^U$ in descending order.
Select user successively by $\rho_k R_k^U$ and calculate the learning efficiency $E_M$ of the FEEL system.
For $M = 1$ to $K$, do
  if $M = 1$, then
    $E_{\max} = E_M$.
  else
    if $E_M < E_{\max}$, then
      break.
    else
      $E_{\max} = E_M$.
End
Calculate the corresponding $\{a_k^*, T^*, \tau_k^{U*}\}$ with $E_{\max}$.
Output the optimal solution $\{a_k^*, T^*, \tau_k^{U*}\}$.
```

5 Simulation Results

In this section, we test the performance of the proposed algorithm by simulation and validate the performance improvement by comparing with other traditional algorithms.

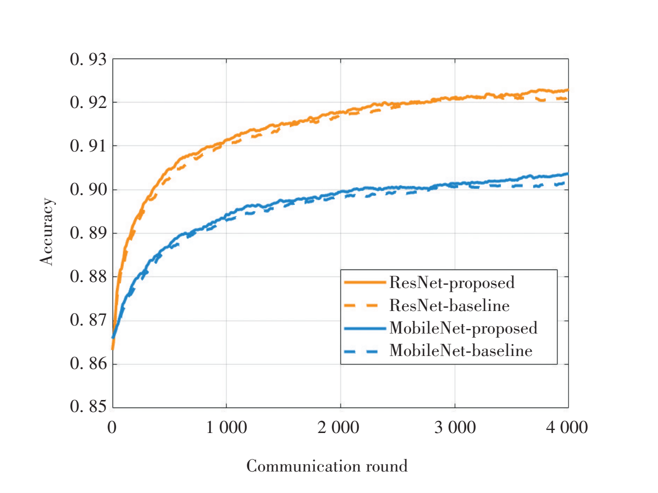
5.1 Simulation Settings

In the FEEL system, K users are stochastically distributed over the coverage of the BS. The coverage area of the BS is a circle with a radius of 500 m. All users are connected with the BS by wireless channels. The channel gains are generated by the pass loss model, 128.1+37.6log(d [km]), while the small-scale fading obeys the Rayleigh distribution with uniform variance. The noise power spectral density is −174 dBm/Hz and the system bandwidth is 5 MHz. The uplink and downlink transmit powers are both 24 dBm.

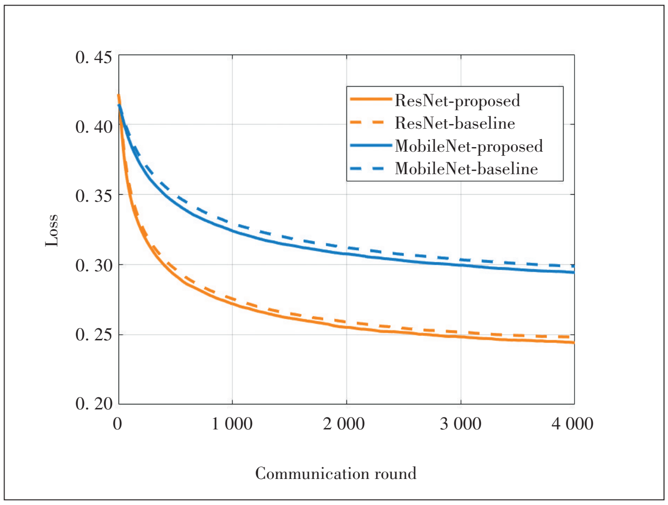
We utilize the dataset CIFAR-10 as the local dataset of all users to train model. The dataset is composed of 60 000 32×32 color images in 10 classes, which includes 50 000 training images and 10 000 test images. We shuffle all training samples first, divide them into K parts equally and then distribute them to all users, respectively. Two common DNN models, MobileNetV2 and ResNet18, are deployed for image classification. Since it is time-consuming to restart training, we utilize the pre-trained model to reduce the model convergence time.

5.2 Tests of Generalization Ability

The generalization ability refers to the adaptability of algorithms to different DNN models. To test the generalization ability of our proposed algorithm, we implement it on the two DNN models as mentioned before when there are K = 14 users in the FEEL system. Meanwhile, we make comparisons with the performance of proposed algorithm and the *baseline algorithm* where all users are selected with equal communication resource allocation. The simulation results of the test accuracy and the global training loss are shown in **Figs. 2** and **3**, respectively. From the figures, the proposed algorithm can achieve a high learning accuracy and a fast convergence rate for different DNN models. The result shows that our proposed algorithm has excellent generalization ability and can be widely implemented in practical systems. Moreover, the performance of our proposed algorithm is similar to that of the *baseline algorithm* with the increase of communication round rather than training time. It is reasonable since our proposed algorithm aims to reduce the communication delay in each communication round, rather than the number of communication rounds. Besides, this result demonstrates that our proposed algorithm can achieve the similar training speed by only selecting partial users in the FEEL system.



▲**Figure 2. The test accuracy versus communication round.**



▲**Figure 3. The global training loss versus communication round.**

5.3 Performance Comparison Among Different Algorithms

In this part, we compare the performance of our proposed algorithm with other conventional algorithms to verify its superiority. The two benchmark algorithms are described as follows.

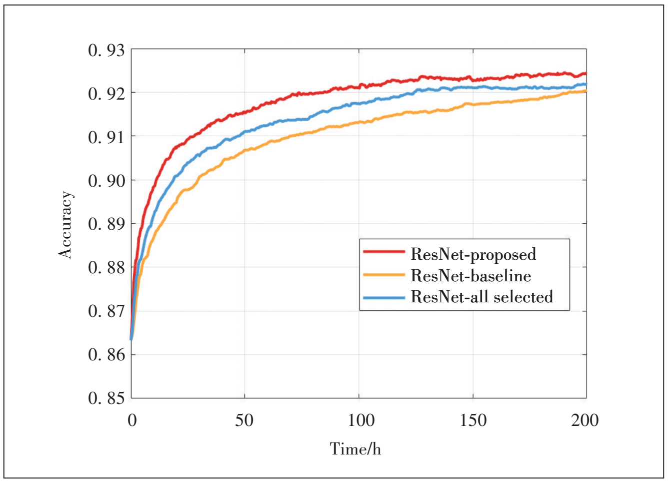
- *Baseline algorithm*: In each communication round, all users in the FEEL system participate in the training task with equal communication resource allocation, i.e., $\tau_k = 1/K, \forall k \in \mathcal{K}$.

- *All selected algorithm*: In each communication round, all users in the FEEL system participate in the training task with optimal communication resource allocation based on Eq. (19).

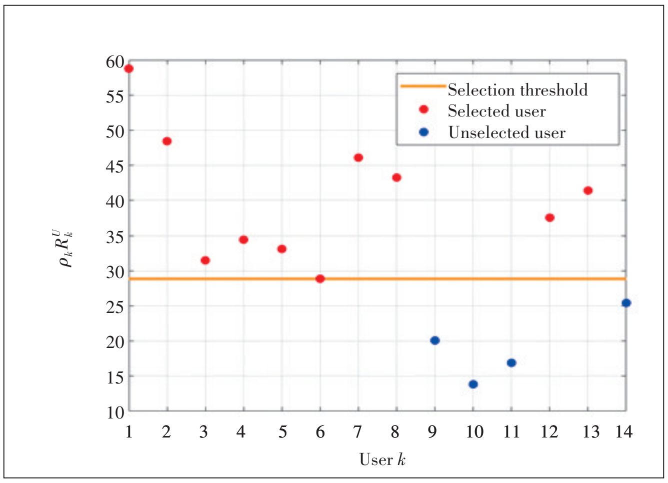
Here we use the pre-trained ResNet18 model to test the performance of the three algorithms in an FEEL system with K = 14 users. The test accuracy versus training time with different algorithms is shown in **Fig. 4**. From the figure, it can be seen that our proposed algorithm achieves the highest test accuracy among all algorithms. The reason is that our proposed algorithm not only selects users based on data importance but also makes the optimal communication resource allocation. By doing so, only users with more important data and higher uplink data rate participate in the training task. Thus, the communication latency is reduced and the global loss decay rate increases, which eventually improves the learning efficiency of the system. The gap between the *baseline algorithm* and the *all selected algorithm* demonstrates the gain obtained by the optimal communication resource allocation. The gap between the *all selected algorithm* and the *proposed algorithm* demonstrates the gain obtained by the optimal user selection. In conclusion, our proposed algorithm accelerates the training task and improves the learning efficiency of the FEEL system by jointly considering user selection and communication resource allocation.

To further verify the applicability and effectiveness of our proposed algorithm, we select one communication round randomly to obtain more simulation results. **Figs. 5** and **6** illustrate the results of user selection and communication resource allocation for our proposed algorithm in the communication round we selected, respectively. From Fig. 5, we can observe that user k is selected only when the product of its data importance and uplink data rate, i.e., $\rho_k R_k^U$, is no less than the selection threshold, which is consistent with Theorem 1. Moreover, in order to clearly present the relationship between the communication resource allocation and the uplink data rate, we plot the corresponding uplink data rate for all users in **Fig. 7**. Combining Fig. 6 with Fig. 7, it can be observed that a selected user with a higher uplink data rate is allocated with less communication resource, which is consistent with Eq. (19).

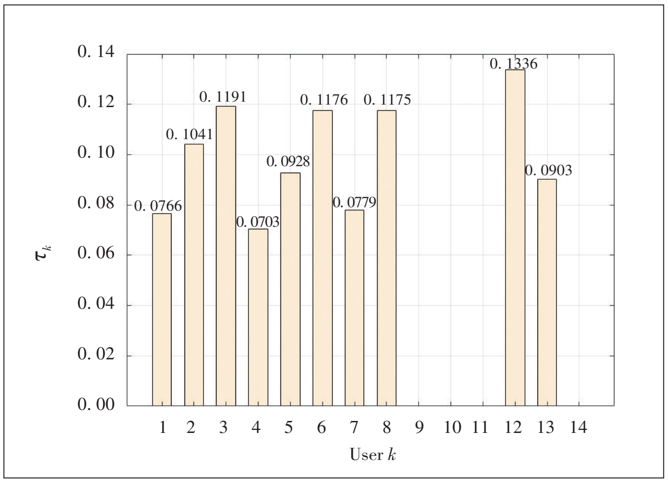
In the end, we further study how the number of users impacts the training performance of the FEEL system. The test accuracy versus training time with different numbers of users is shown in **Fig. 8**. From the figure, it can be seen that our proposed algorithm achieves the highest system learning efficiency when K = 6. The reasons can be explained as follows. The number of time slot allocated to the selected user is large when the number of users is small. Consequently, the communication latency greatly reduces, and the learning efficiency of the FEEL system significantly improves in this scenario. Moreover, the number of selected users is limited by the scarce wireless communication resource when the number of users is too large. Therefore, the learning efficiency of the FEEL system does not improve with user number when too many users in the system.



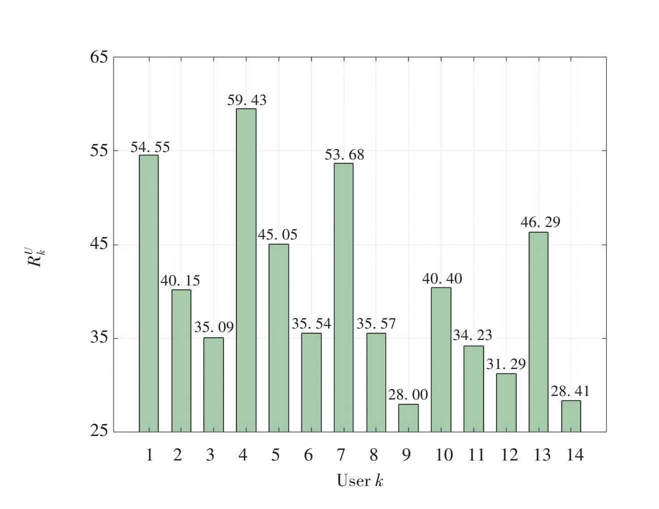
▲Figure 4. The test accuracy versus training time with different algorithms.



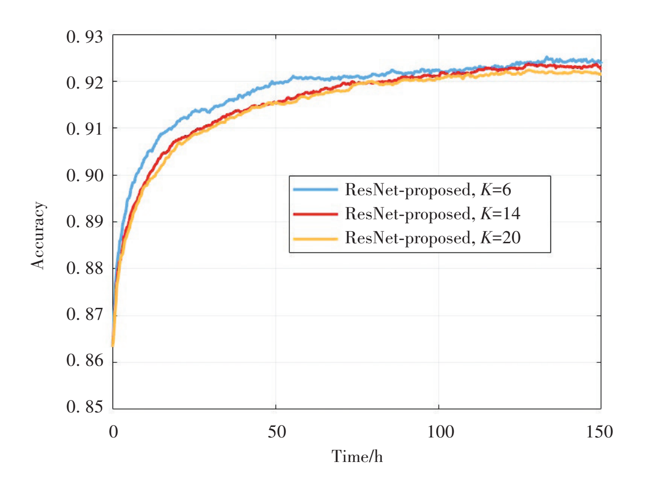
▲Figure 5. User selection for the proposed algorithm.



▲ Figure 6. Communication resource allocation for the proposed algorithm.



▲**Figure 7. Corresponding uplink data rate.**



▲**Figure 8. Test accuracy versus training time with different numbers of users.**

6 Conclusions

In this paper, we aim to accelerate the training task and improve the learning efficiency of the FEEL system by proposing an optimal user selection policy based on data importance and CSI. After analyzing the data importance of users and the end-to-end latency of the FEEL system, we formulate an optimization problem to maximize the learning efficiency of the FEEL system. By problem transformation and relaxation, we first develop the optimal user selection policy. Based on this, the optimal communication resource allocation is developed in closed-form. We further develop a polynomial-time algorithm to solve this mixed-integer programming problem and prove its optimality. Finally, the simulation results show that our proposed algorithm has strong generalization ability and can significantly improve the learning efficiency of the FEEL system.

Our work has demonstrated that the learning efficiency of the FEEL system can be further improved by user selection based on data importance and wireless resource allocation. However, some assumptions have been made to gain insightful results. In the future, we will make further investigation to better fit the practical systems. First, we have assumed that there is no inter-cell interference in the uplink. In the future, the FEEL system with inter-cell interference deserves further investigation. Second, the local gradient received by the edge server may contain data errors, which may affect the training performance of the FEEL system. Therefore, our future work can further study the impact of those errors. Last but not the least, it is meaningful to extend our proposed algorithm to the FEEL system, where orthogonal frequency-division multiple access (OFDMA) is adopted for data transmission.

Appendix A

Proof of Theorem 1

We apply the Lagrangian method to obtain the optimal solution to problem $\mathcal{P}2$ with fixed T since it is a convex optimization problem. The Lagrangian function is defined as

$$L = -\frac{\sum_{k=1}^{K} a_k \rho_k}{T} + \lambda\left(\sum_{k=1}^{K} \frac{a_k V}{R_k^U} - T + T^C\right), \tag{20}$$

where λ is the Lagrange multiplier related with the constraint (17b). By applying the Karush-Kuhn-Tucker (KKT) conditions and simple calculation, we can draw the following necessary and sufficient conditions, as

$$\frac{\partial L}{\partial a_k^*} = -\frac{\rho_k}{T} + \lambda^* \frac{V}{R_k^U} \begin{cases} \geqslant 0, & a_k^* = 0, \\ = 0, & 0 \leqslant a_k^* \leqslant 1, \ \forall k \in \mathcal{K}, \\ \leqslant 0, & a_k^* = 1, \end{cases} \tag{21}$$

$$\lambda^* \left(\sum_{k=1}^{K} \frac{V a_k^*}{R_k^U} + T^C - T\right) = 0,\ \lambda^* \geqslant 0. \tag{22}$$

With simple mathematical calculation, we can derive the optimal user selection policy as shown in Theorem 1, which ends the proof.

Appendix B

Proof of Theorem 2

According to Theorem 1, users are selected by the descending order of $\rho_k R_k^U$. Hence, we can assume that $a_k = 1$ when $k = 1, 2, \ldots, M$ and $a_k = 0$ when $k = M + 2, M + 3, \ldots, K$. Moreover, it is not clear whether $a_{M+1} = 0$ or $a_{M+1} = 1$. Then, we denote $E^{(1)}$ as the objective function of problem $\mathcal{P}2$, which can be expressed as

$$E^{(1)} = \frac{\sum_{k=1}^{M}\rho_k + a_{M+1}\rho_{M+1}}{\sum_{k=1}^{M}\frac{V}{R_k^U} + a_{M+1}\frac{V}{R_{M+1}^U} + T^C}. \tag{23}$$

So the derivative of $E^{(1)}$ with respect to a_{M+1} is given by

$$\frac{\partial E^{(1)}}{\partial a_{M+1}} = \frac{\rho_{M+1}\left(\sum_{k=1}^{M}\frac{V}{R_k^U} + T^C\right) - \frac{V}{R_{M+1}^U}\sum_{k=1}^{M}\rho_k}{\left(\sum_{k=1}^{M}\frac{V}{R_k^U} + a_{M+1}\frac{V}{R_{M+1}^U} + T^C\right)^2}. \tag{24}$$

It shows that the sign of derivative is consistent with the sign of the numerator of Eq. (24). However, the value of the numerator of Eq. (24) is independent of a_{M+1}. Therefore, $E^{(1)}$ is monotone when $a_{M+1} \in [0,1]$. That is, the maximum value of $E^{(1)}$ must be obtained either when $a_{M+1} = 0$ or when $a_{M+1} = 1$. In conclusion, the optimal solution of a_k to $\mathcal{P}2$ must be an integer solution. Hence, this solution must be the feasible solution to problem $\mathcal{P}1$ as well. Moreover, after relaxation, the maximum value of the objective function is non-decreasing. Thus, the optimal solutions to problem $\mathcal{P}2$ and $\mathcal{P}1$ are exactly the same, which ends the proof.

Appendix C

Proof of Theorem 3

According to Theorem 2, we know that the optimal solutions to problem $\mathcal{P}2$ and $\mathcal{P}1$ are exactly the same. Thus, we only consider the integer solutions here. When no user is selected, the learning efficiency is zero obviously. The learning efficiency must increase first with the number of selected users. In other words, at least one user is selected. Then we consider the following condition.

Denote $T_M = \sum_{k=1}^{M} V/R_k^U + T^C, M \in \{1,2,\dots,K\}$ as the system latency when M users are selected. Assume that the following formulas exist

$$E_M - E_{M-1} = \frac{\rho_M\sum_{k=1}^{M-1}\frac{V}{R_k^U} + T^C\rho_M - \frac{V}{R_M^U}\sum_{k=1}^{M-1}\rho_k}{\left(\sum_{k=1}^{M}\frac{V}{R_k^U} + T^C\right)\left(\sum_{k=1}^{M-1}\frac{V}{R_k^U} + T^C\right)} > 0, \tag{25}$$

$$E_{M+1} - E_M = \frac{\rho_{M+1}\sum_{k=1}^{M}\frac{V}{R_k^U} + T^C\rho_{M+1} - \frac{V}{R_{M+1}^U}\sum_{k=1}^{M}\rho_k}{\left(\sum_{k=1}^{M+1}\frac{V}{R_k^U} + T^C\right)\left(\sum_{k=1}^{M}\frac{V}{R_k^U} + T^C\right)} < 0. \tag{26}$$

From Eqs. (25) and (26), we can obtain the following inequalities.

$$\rho_M R_M^U\left(\sum_{k=1}^{M-1}\frac{V}{R_k^U} + T^C\right) > V\sum_{k=1}^{M-1}\rho_k, \tag{27}$$

$$\rho_{M+1} R_{M+1}^U\left(\sum_{k=1}^{M}\frac{V}{R_k^U} + T^C\right) < V\sum_{k=1}^{M}\rho_k. \tag{28}$$

According to Eq. (27), we can derive the recurrence formula as

$$\begin{aligned}&\rho_{M-1}R_{M-1}^U\left(\sum_{k=1}^{M-2}\frac{V}{R_k^U} + T^C\right) - V\sum_{k=1}^{M-2}\rho_k = \\ &\rho_{M-1}R_{M-1}\left(\sum_{k=1}^{M-1}\frac{V}{R_k^U} + T^C\right) - V\sum_{k=1}^{M-1}\rho_k > \\ &\rho_M R_M^U\left(\sum_{k=1}^{M-1}\frac{V}{R_k^U} + T^C\right) - V\sum_{k=1}^{M-1}\rho_k > 0,\end{aligned} \tag{29}$$

which implies $E_{M-2} < E_{M-1}$. Then we can obtain the conclusion recursively, as

$$E_1 < E_2 < \dots < E_M. \tag{30}$$

Similar to the above analysis, we have the following conclusion, as

$$E_1 < E_2 < \dots < E_M > E_{M+1} > E_{M+2} > \dots > E_K. \tag{31}$$

Based on the above analysis, E_M first increases and then decreases with M, which ends the proof.

References

[1] ZHOU Z, CHEN X, LI E, et al. Edge intelligence: paving the last mile of artificial intelligence with edge computing [J]. Proceedings of the IEEE, 2019, 107 (8): 1738 – 1762. DOI: 10.1109/jproc.2019.2918951

[2] ZHU G, LIU D Z, DU Y Q, et al. Towards an intelligent edge: wireless communication meets machine learning [EB/OL]. (2018-09-02)[2019-12-05]. https://arxiv.org/abs/1809.00343

[3] MCMAHAN 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. Fort Lauderdale, USA, 2017: 1273 – 1282.

[4] ETSI. Mobile edge computing-introductory technical white paper [R]. 2014

[5] ZHU G X, WANG Y, HUANG K B. Broadband analog aggregation for low-latency federated edge learning (extended version) [EB/OL]. (2018-10-30)[2019-01-16]. https://arxiv.org/abs/1812.11494

[6] NISHIO T, YONETANI R. Client selection for federated learning with heterogeneous resources in mobile edge [C]//IEEE International Conference on Communications (ICC). Shanghai, China, 2019: 18852422. DOI: 10.1109/icc.2019.8761315

[7] ZENG Q S, DU Y Q, LEUNG K K, et al. Energy-efficient radio resource allocation for federated edge learning [EB/OL]. (2019-07-13)[2019-12-20]. https://arxiv.org/abs/1907.06040

[8] YANG Z H, CHEN M Z, SAAD W, et al. Energy efficient federated learning over wireless communication networks [EB/OL]. (2019-11-06)[2019-12-10]. https://arxiv.org/abs/1911.02417

[9] YANG H H, LIU Z Z, QUEK T Q S, et al. Scheduling policies for federated learning in wireless networks [J]. IEEE transactions on communications, 2020, 68(1): 317 – 333. DOI: 10.1109/tcomm.2019.2944169

[10] YANG H H, ARAFA A, QUEK T Q S, et al. Age-based scheduling policy for federated learning in mobile edge networks [C]//IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain, 2020. DOI: 10.1109/icassp40776.2020.9053740

[11] CHEN M Z, YANG Z H, SAAD W, et al. A joint learning and communications framework for federated learning over wireless networks [EB/OL]. (2019-09-17)[2020-01-10]. https://arxiv.org/abs/1909.07972

[12] CHEN M Z, POOR H V, SAAD W, et al. Convergence time optimization for federated learning over wireless networks [EB/OL]. (2020-01-22)[2020-03-25]. https://arxiv.org/abs/2001.07845

[13] KATHAROPOULOS A, FLEURET F. Not all samples are created equal: deep learning with importance sampling [EB/OL]. (2018-03-02)[2019-10-28]. https://arxiv.org/abs/1803.00942

[14] LIU D Z, ZHU G X, ZHANG J, et al. Wireless data acquisition for edge learning: data-importance aware retransmission [EB/OL]. (2018-10-05)[2019-03-19]. https://arxiv.org/abs/1812.02030

[15] LIU D Z, ZHU G X, ZHANG J, et al. Data-importance aware user scheduling for communication-efficient edge machine learning [EB/OL]. (2019-03-19)[2019-10-05]. https://arxiv.org/abs/1910.02214

[16] ETSI. LTE; evolved universal terrestrial radio access (E-UTRA); physical channels and modulation (3GPP TS 36.211 version 15.6.0 release 15): ETSI TS 136 211 V15.6.0 [S]. ETSI, 2019

[17] LIN Y J, HAN S, MAO H Z, et al. Deep gradient compression: reducing the communication bandwidth for distributed training [EB/OL]. (2017-10-05)[2018-02- 05]. https://arxiv.org/abs/1712.01887

[18] REN J K, YU G D, CAI Y L, et al. Latency optimization for resource allocation in mobile-edge computation offloading [J]. IEEE transactions on wireless communications, 2018, 17(8): 5506 – 5519. DOI: 10.1109/twc.2018.2845360

[19] CHEN T Y, GIANNAKIS G B, SUN T, et al. LAG: lazily aggregated gradient for communication-efficient distributed learning [EB/OL]. (2018-05-25)[2019-12-02]. https://arxiv.org/abs/1805.09965

[20] REN J K, YU G D, and DING G Y. Accelerating DNN training in wireless federated edge learning system [EB/OL]. (2019-05-23)[2020-03-28]. https://arxiv.org/abs/1905.09712

Biographies

JIANG Zhihui (zhihui.jiang@zju.edu.cn) received the B.E. degree in information engineering from Zhejiang University, China in 2020, where she is currently pursuing the master's degree with the College of Information Science and Electronic Engineering. Her research interests mainly include federated learning and edge learning.

HE Yinghui received the B.E. degree in information engineering from Zhejiang University, China in 2018, where he is currently pursuing the master's degree with the College of Information Science and Electronic Engineering. His research interests mainly include mobile edge computing and device-to-device communications.

YU Guanding received the B.E. and Ph.D. degrees in communication engineering from Zhejiang University, China in 2001 and 2006, respectively. He joined Zhejiang University, in 2006, where he is currently a full professor with the College of Information and Electronic Engineering. From 2013 to 2015, he was a visiting professor with the School of Electrical and Computer Engineering, Georgia Institute of Technology, USA. His research interests include 5G communications and networks, mobile edge computing, and machine learning for wireless networks. Dr. YU received the 2016 IEEE ComSoc Asia-Pacific Outstanding Young Researcher Award. He regularly chairs the technical program committee boards of prominent IEEE conferences, such as ICC, GLOBECOM, and VTC. He also serves as a symposium co-chair for IEEE Globecom 2019 and the track chair for IEEE VTC 2019' Fall. He has served as a guest editor for the *IEEE Communications Magazine* special issue on full-duplex communications, an editor for the *IEEE Journal on Selected Areas* in Communications Series on green communications and networking, a leading guest editor for the *IEEE Wireless Communications Magazine* special issue on LTE in unlicensed spectrum, and an editor for the *IEEE Access*. He serves as an editor for the *IEEE Transactions on Green Communications and Networking* and the *IEEE Wireless Communications Letters*.

⬅ **From Page 01**

future directions. The third paper "Joint User Selection and Resource Allocation for Fast Federated Edge Learning" by JIANG et al. presents a new policy for joint user selection and communication resource allocation to accelerate the training task and improve the learning efficiency.

Edge learning includes both edge training and edge inference. Due to the stringent latency requirements, edge inference is particularly bottlenecked by the limited computation and communication resources at the network edge. The fourth paper "Communication-Efficient Edge AI Inference over Wireless Networks" by YANG et al. identifies two communication-efficient architectures for edge inference, namely, on-device distributed inference and in-edge cooperative inference, thereby achieving low latency and high energy efficiency. The fifth paper "Knowledge Distillation for Mobile Edge Computation Offloading" by CHEN et al. introduces a new computation offloading framework based on deep imitation learning and knowledge distillation that assists end devices to quickly make fine-grained offloading decisions so as to minimize the end-to-end task inference latency in MEC networks. By considering edge inference in MEC-enabled UAV systems, the last paper "Joint Placement and Resource Allocation for UAV-Assisted Mobile Edge Computing Networks with URLLC" by ZHANG et al. jointly optimizes the UAV's placement location and transmitting power to facilitate ultra-reliable and low-latency round-trip communication from sensors to UAV servers to actuators.

We hope that the aforementioned six papers published in this special issue stimulate new ideas and innovations from both the academia and industry to advance this exciting area of edge learning.