Optimal Resource Allocation for U-Shaped Parallel Split Learning

Song Lyu*, Zheng Lin*, Guanqiao Qu*, Xianhao Chen*, Xiaoxia Huang†, and Pan Li‡

*The Department of Electrical and Electronic Engineering, The University of Hong Kong, Pok Fu Lam, Hong Kong, China
†School of Electronics and Communication Engineering, Sun Yat-sen University, Shenzhen 510275, China
‡The Department of Electrical, Computer, and System Engineering, Case Western Reserve University,
Cleveland, OH 44106 USA.

Abstract—Split learning (SL) has emerged as a promising approach for model training without revealing the raw data samples from the data owners. However, traditional SL inevitably leaks label privacy as the tail model (with the last layers) should be placed on the server. To overcome this limitation, one promising solution is to utilize U-shaped architecture to leave both early layers and last layers on the user side. In this paper, we develop a novel parallel U-shaped split learning and devise the optimal resource optimization scheme to improve the performance of edge networks. In the proposed framework, multiple users communicate with an edge server for SL. We analyze the end-to-end delay of each client during the training process and design an efficient resource allocation algorithm, called LSCRA, which finds the optimal computing resource allocation and split layers. Our experimental results show the effectiveness of LSCRA and that U-shaped parallel split learning can achieve a similar performance with other SL baselines while preserving label privacy.

Index Terms—U-shaped network, split learning, label privacy, resource allocation, 5G/6G edge networks.

I. INTRODUCTION

Traditional centralized learning incurs excessive bandwidth consumption and communication latency while violating data privacy. To address this issue, edge learning, which trains models at the network edge, has emerged as a promising paradigm in 5G and beyond [1]–[4]. In this respect, federated edge learning (FEEL) [5], [6] has been shown as an effective approach that enables end devices to train models on their own devices and then aggregate the models at an edge server, thereby eliminating the need for access the raw data.

However, FEEL faces significant challenges due to the extensive client-side computing workload. For massive resource-constrained IoT devices, the limited computing power may hinder their ability to perform model training and upload large models [7]. To address these challenges, split learning (SL) [8] has emerged as an effective technique. SL splits the model into two parts: the front sub-model (head model) trained by a client and the remaining sub-model (tail model) trained by a server [9]. As a result, SL significantly relieves clients' computing burden by allowing a server to take over the major workload while remaining raw data on the client side [10]–[12].

There exist several popular SL approaches. Vanilla SL has limited scalability due to its sequential training manner [8], where the model training can be shifted to the next client only when the previous client completes training. To parallize SL, parallel split learning (PSL) [13] enables parallel processing across the server and multiple connected clients. Furthermore, split federated learning (SFL) [14] integrates federated learning (FL) into SL to allow parallel training. U-shaped split federated learning (U-SFL) [15] combines the U-shaped architecture with the SFL framework to eliminate label sharing. Compared to PSL, the major change in SFL lies in the averaging of the client-side sub-model after its backpropagation process, following the spirit of FL, yet incurring additional communication overhead due to model exchange. The comparison of these approaches is summarized in Table I.

By preserving users' raw data, SL is often considered in privacy-sensitive applications [16]. Nevertheless, despite the preservation of input data, the sharing of label can be a serious privacy concerns in SL, as clients have to provide the corresponding labels to help the server to calculate the loss. In some applications, label privacy is an important concern, particularly in healthcare, finance, and other sensitive domains. For example, the input data can be users' bio information/activities, and the label is the disease or health status of this user. In this case, the label is also highly sensitive and should not be shared with the server.

To address the label privacy issue, U-shaped configurations has been proposed for SL to eliminate the need for label exchange [8]. In the U-shaped SL architecture, the entire DNN is divided into three submodels: the head, body, and tail models. The head and tail models are obtained on the client side, while the body model is trained on the edge server side. This architecture effectively resolves the label privacy concern, as the output layer and the labels are retained on the client side. Although U-shaped SL has been studied under various contexts, such as medical applications [16], to our best knowledge, very few efforts have been made to integrate U-shaped SL into the mobile edge.

In this paper, we investigate U-Shaped Parallel Split Learning (U-PSL) under the mobile edge computing framework. This framework parallelizes the vanilla U-shaped SL by enabling multiple clients to train with a server simultaneously. Furthermore, we develop the joint model split and resource allocation problem tailored for U-shaped SL, called LSCRA. By formulating

TABLE I:

The Comparison of FL, SL, SFL, PSL, U-SFL, and U-PSL Frameworks

Learning framework	FL	SL	SFL	PSL	U-SFL	U-PSL
Computation offloading	No	Yes	Yes	Yes	Yes	Yes
Parallel computing	Yes	No	Yes	Yes	Yes	Yes
Access to raw data	No	No	No	No	No	No
Model exchange	Yes	No	Yes	No	Yes	No
Label sharing	No	Yes	Yes	Yes	No	No

Fig. 1: The illustration of U-PSL over wireless networks.

the per-round training latency, we obtain the optimal server computing resource allocation and layer splitting strategy to address the communication and computing challenges associated with U-shaped networks, resulting in a significant reduction in training latency. Through experiments, it is found that U-PSL achieves effective label privacy protection while achieving similar or even slightly shorter latency compared to other SL benchmarks, making it a promising solution for SL in privacy-sensitive and resource-constrained wireless networks.

Our contributions are summarised as follows:

- We propose U-PSL, an advanced privacy-enhancing training framework, which eliminates the need for raw data sharing and label sharing in SL.
- We design an optimal joint computing resource allocation and layer splitting scheme to minimize per-round latency.
- We conduct simulations to demonstrate the effectiveness of the U-PSL framework. Our simulations show the effectiveness of the resource allocation scheme, revealing that the framework achieves test accuracy comparable to other SL approaches while preserving label privacy.

II. SYSTEM MODEL AND U-PSL FRAMEWORK

This section presents the U-PSL framework, which is illustrated in Figure 1. We begin by describing a scenario of the U-PSL framework in wireless networks. Subsequently, we provide a detailed explanation of the five main steps in the U-PSL workflow. Through this section, we aim to provide an overview of the U-PSL framework and its step-by-step training procedure. Furthermore, since a shorter training time not only enables timely model usage but also reduces bandwidth and computing resource occupation, we will analyze and optimize the end-to-end latency. For the convenience of readers, we summarize the important notations in Table II.

Architecture: U-PSL comprises an edge server and multiple clients. On the client side, we assume that each client has an end device with computing capabilities, enabling it to execute forward propagation (FP) and backpropagation (BP) for the client-side models. Let $\mathcal{U} = \{1, 2, ..., N\}$ denote the set of clients, where N is the number of participating clients. The local dataset D_n owned by client $n \in \mathcal{U}$ is represented as $D_n = \{X_n, Y_n\}$, where X_n denotes the n-th client's training dataset and Y_n is the set of the corresponding labels. ρ_j and ω_j denote the computing workload of FP and BP for the first j layers, respectively, ψ_j represents the activation size at cut layer j in the model, and L denotes the total number of model layers.

U-PSL Workflow: Figure 2 illustrates the main workflow of U-PSL, which consists of five training steps:

1) Head model FP & activations transmission: At the beginning of model training, the server initializes the global model and partitions it into three submodels W_{head}, W_{body}, and W_{tail}. $\mu_1^j = \{0, 1\}$ and $\mu_2^j = \{0, 1\}$ indicates the two split layers between head models and body models, and body models and tail models. $\mu_1^j = 1$ indicates that layer j is the first cut layer, and $\mu_2^j = 1$ indicates that layer j is the second cut layer. At the beginning of each round, each client randomly draws a mini-batch β_n (generally, the size can be proportional to the size of X_n) to perform the FP process of the head model in parallel. For simplicity, we focus on client n to illustrate the operations on the client side. Let $\varphi_1^F(\overline{\mu_1^j})$ denote the computing workload of the head model's FP process for one data sample, which is given by:

$$\varphi_1^F(\overline{\mu_1^j}) = \sum_{j=1}^{L} \mu_1^j \rho_j. \tag{1}$$

After completing the head model FP process, the first cut layer generates activations that will be taken as the input of the body model on the server. Then, the client transmits the activations to the server over a wireless channel. The data size of the activations $\Gamma(\overline{\mu_1^j})$ can be expressed as:

$$\Gamma(\overline{\mu_1^j}) = \sum_{j=1}^{L} \mu_1^j \psi_j. \tag{2}$$

Therefore, the latency of step 1 for client n can be denoted as:

$$t_{1,n}^c = \frac{\beta_n \varphi_1^F(\overline{\mu_1^j}) K_c}{f_n} + \frac{\beta_n \Gamma(\overline{\mu_1^j})}{R_n^\uparrow}, \tag{3}$$

where f_n is the computing capability of client n, K_c is the computing intensity of client, and $R_n^\uparrow$ is the upload data rate. We consider a static network where the average data rate does not change, and therefore $R_n^\uparrow$ is a constant value. The mobility scenarios can be left for the future research [17]–[20].

2) Body model FP & activations transmission: When the server receives the activations from clients, the body model starts its FP process. $\varphi_s^F(\overline{\mu_1^j}, \overline{\mu_2^j})$ denotes the computation workload

TABLE II: Frequently Used Notations

Notation	Interpretation
$\mathcal{U}$	The set of clients
D_n	Local dataset of client n
β_n	mini-batch size draw from client n's local dataset
$R_n^\uparrow/R_n^\downarrow$	Upload/download data rate of client n
f_n	The computing capability of client n
$f_{s,n}$	The server-side computing resource allocated for client n
K_s/K_c	The computing intensity of server/client n
L	The total number of model layers in CNN
ρ_j	The computation workload (in CPU cycles) of FP for the first j layers
ω_j	The computation workload (in CPU cycles) of BP for the first j layers
ψ_j	The size of activations (or activations' gradients) of the cut layer j
F_s	Maximum computing capability of the server

Fig. 2: U-PSL Framework.

of the body model's FP process for one data sample, which can be described as:

$$\varphi_s^F(\overline{\mu_1^j}, \overline{\mu_2^j}) = \sum_{j=1}^{L} \mu_2^j \rho_j - \sum_{j=1}^{L} \mu_1^j \rho_j. \qquad (4)$$

After the completion of each mini-batch, the second cut layer generates activations, the size of which can be expressed as:

$$\Gamma(\overline{\mu_2^j}) = \sum_{j=1}^{L} \mu_2^j \psi_j. \qquad (5)$$

Therefore, the latency of step 2 for client n can be denoted as:

$$t_{2,n}^s = \frac{\beta_n \varphi_s^F(\overline{\mu_1^j}, \overline{\mu_2^j}) K_s}{f_{s,n}} + \frac{\beta_n \Gamma(\overline{\mu_2^j})}{R_n^\downarrow}. \qquad (6)$$

where $f_{s,n}$ is the server computing resource allocation of client n, K_s is the computing intensity of server, and $R_n^\downarrow$ is the download data rate.

3) Tail model FP and BP & activations' gradients transmission: At this stage, the client performs the rest FP process of the tail model to calculate the loss and then conducts the BP process. Let $\varphi_2^F(\overline{\mu_2^j})$ and $\varphi_2^B(\overline{\mu_2^j})$ represent the computation workload of the tail model's FP and BP process, respectively, which can be described as:

$$\varphi_2^F(\overline{\mu_2^j}) = \rho_L - \sum_{j=1}^{L} \mu_2^j \rho_j, \qquad (7)$$

$$\varphi_2^B(\overline{\mu_2^j}) = \omega_L - \sum_{j=1}^{L} \mu_2^j \omega_j. \qquad (8)$$

After finishing the BP process, each client sends activations' gradients back to the server. Given the size of the activations' gradients at the second layer $\Gamma(\overline{\mu_2^j})$ in (5), the latency of step 3 for client n can be denoted as:

$$t_{3,n}^c = \frac{\beta_n K_c(\varphi_2^F(\overline{\mu_2^j}) + \varphi_2^B(\overline{\mu_2^j}))}{f_n} + \frac{\beta_n \Gamma(\overline{\mu_2^j})}{R_n^\uparrow}. \qquad (9)$$

4) Body model BP & activations' gradients transmission: After receiving activations' gradients, the server performs its BP process. Let $\varphi_s^B(\overline{\mu_1^j}, \overline{\mu_2^j})$ denotes the computation workload of the body model's BP process, which can be described as:

$$\varphi_s^B(\overline{\mu_1^j}, \overline{\mu_2^j}) = \sum_{j=1}^{L} \mu_2^j \omega_j - \sum_{j=1}^{L} \mu_1^j \omega_j. \qquad (10)$$

When the body model's BP process is completed, activations' gradients at the first cut layer will be transmitted to the corresponding clients. The size of activations' gradients is $\Gamma(\overline{\mu_1^j})$ in (2), and therefore the latency of step 4 for client n can be denoted as:

$$t_{4,n}^s = \frac{\beta_n \varphi_s^B(\overline{\mu_1^j}, \overline{\mu_2^j}) K_s}{f_{s,n}} + \frac{\beta_n \Gamma(\overline{\mu_1^j})}{R_n^\downarrow}. \qquad (11)$$

5) Head model BP: In this stage, the client only needs to complete the rest BP process of the head model. $\varphi_1^B(\overline{\mu_1^j})$ denotes the computation workload of the head model's BP process, which can be described as:

$$\varphi_1^B(\overline{\mu_1^j}) = \sum_{j=1}^{L} \mu_1^j \omega_j. \qquad (12)$$

Therefore, the latency of step 5 for client n can be denoted as:

$$t_{5,n}^c = \frac{\beta_n \varphi_1^B(\overline{\mu_1^j}) K_c}{f_{s,n}}. \qquad (13)$$

After the aforementioned steps, each sub-model updates the model parameters according to the gradients. Note that, for the body model, the server can make updates based on the averaged gradients across the clients. The per-round training latency corresponding to client n can be denoted as:

$$T_n = t_{1,n}^c + t_{2,n}^s + t_{3,n}^c + t_{4,n}^s + t_{5,n}^c. \qquad (14)$$

Let $T(f, \mu_1, \mu_2)$ denote the per-round training time. Since the aforementioned training is executed in parallel, $T(f, \mu_1, \mu_2)$ is equal to the maximum T_n, i.e.,

$$T(f, \mu_1, \mu_2) = \max_{n \in \mathcal{U}} T_n. \qquad (15)$$

III. Problem Formulation and Solution Approach

As mentioned earlier, the total latency of one training round for a client is formulated. Apparently, inappropriate server computing resource allocation can lead to significant increases in training time. Additionally, the selection of cut layers also affects the overall training and communication latency. Considering these factors, we formulate the following optimization problem to minimize the per-round latency:

$$\mathcal{P}1 : \min_{\mathbf{f}, \boldsymbol{\mu_1}, \boldsymbol{\mu_2}} T(\mathbf{f}, \boldsymbol{\mu_1}, \boldsymbol{\mu_2}) \tag{16}$$

$$\text{s.t. C1}: \sum_{j'=1}^{j} \mu_2^{j'} \leq \sum_{j'=1}^{j} \mu_1^{j'}, \forall j \in \{1, ..., L\},$$

$$\text{C2}: \mu_2^j \in \{0,1\}, \mu_1^j \in \{0,1\}, \forall j \in \{1, ..., L\},$$

$$\text{C3}: \sum_{j=1}^{L} \mu_1^j = 1, \sum_{j=1}^{L} \mu_2^j = 1,$$

$$\text{C4}: f_{s,n} \geq 0, \forall n \in \mathcal{U},$$

$$\text{C5}: \sum_{n=1}^{N} f_{s,n} \leq F_s.$$

where C1 ensures that the index of the second split layer is greater than the index of the first split layer. To solve $\mathcal{P}1$, we first consider the subproblem involving computing resource allocation:

$$\mathcal{P}2 : \min_{\mathbf{f}} T(\mathbf{f}) \tag{17}$$

$$\text{s.t. C4}: f_{s,n} \geq 0, \forall n \in \mathcal{U},$$

$$\text{C5}: \sum_{n=1}^{N} f_{s,n} \leq F_s.$$

We have the following lemmas for $\mathcal{P}2$.

Lemma 1. *The optimal* $\mathbf{f}$ *for* $\mathcal{P}2$ *is obtained when* $T_1 = \cdots = T_n$.

Proof. Let $f_{s,n} = f_{s,n}^*$ be the solution that minimizes the objective while satisfying $T_1 = \cdots = T_n$. Assume $T_1 = \cdots = T_n = \overline{T}$ and therefore $T(\mathbf{f}) = \max_n T_n = \overline{T}$ in this case. It can be shown that $\sum_{n=1}^{N} f_{s,n}^* = F_s$. Otherwise, if $\sum_{n=1}^{N} f_{s,n}^* < F_s$, the remaining resources can be evenly allocated to every $f_{s,n}^*$, thereby reducing the objective. Supposing that there is $T_m > \overline{T}$, we have $T(\mathbf{f}) \geq T_m > \overline{T}$. On the other hand, if there is $T_m < \overline{T}$, we have $f_{s,m} > f_{s,m}^*$. Thus, there must be $f_{s,n} < f_{s,n}^*$ since $\sum_{n=1}^{N} f_{s,n}^* = F_s$. Hence, we have $T_n > \overline{T}$, leading to $T(\mathbf{f}) \geq T_n > \overline{T}$. Therefore, only when $T_1 = \cdots = T_n = \overline{T}$, the optimal resource allocation can be obtained. The proof is completed. $\square$

Lemma 2. *The* k*-th client with the maximum allocated computing resource* $f_{s,k}$ *should satisfy the equation:*

$$\sum_{n=1}^{N} \frac{\varepsilon_n f_{s,k}}{\varepsilon_k + f_{s,k}(T_k^{local} - T_n^{local})} = F_s. \tag{18}$$

Proof. When μ_1^j and μ_2^j are fixed, each training epoch latency can be described as $T_n = T_n^{local} + \frac{\varepsilon_n}{f_{s,n}}$, where $\varepsilon_n = \beta_n K_s(\sum_{j=1}^{L} \mu_2^j \rho_j + \sum_{j=1}^{L} \mu_2^j \omega_j - \sum_{j=1}^{L} \mu_1^j \rho_j - \sum_{j=1}^{L} \mu_1^j \omega_j)$ denotes the server-side computing workload, and T_n^{local} is a constant representing client's local computing and communication latency.

For client set $\mathcal{U}$, by enforcing $T_1 = \cdots = T_k = T_n = \overline{T}, k \in \mathcal{U}$, the equation can be given as

$$T_n^{local} + \frac{\varepsilon_n}{f_{s,n}} = T_k^{local} + \frac{\varepsilon_k}{f_{s,k}}, \forall k, n \in \mathcal{U}. \tag{19}$$

Therefore, to achieve equal per-round training time, we have

$$f_{s,n} = \frac{\varepsilon_n f_{s,k}}{\varepsilon_k + f_{s,k}\left(T_k^{local} - T_n^{local}\right)}. \tag{20}$$

To satisfy C4 in $\mathcal{P}2$, the selected k-th client should be the one with the maximum T_n^{local} to ensure $f_{s,n}$ is nonnegative. Besides, as discussed in Lemma 1, $\sum_{n=1}^{N} f_{s,n} = F_s$ holds for the optimal solution. By considering (20), we have

$$\sum_{n=1}^{N} \frac{\varepsilon_n f_{s,k}}{\varepsilon_k + f_{s,k}(T_k^{local} - T_n^{local})} = F_s. \tag{21}$$

$\square$

We observe that Eq. (18) exhibits a monotonically increasing behavior with respect to $f_{s,k}$. Taking advantage of this property, we can employ a bisection procedure to efficiently find $f_{s,k}$ from (18). Then, the optimal $f_{s,n}$ for other clients can be directly obtained from (20).

After obtaining the optimal server computing resource allocation scheme, the remaining task involves making split-layer decisions. This subproblem can be formulated as:

$$\mathcal{P}3 : \min_{\boldsymbol{\mu_1}, \boldsymbol{\mu_2}} T(\boldsymbol{\mu_1}, \boldsymbol{\mu_2}) \tag{22}$$

$$\text{s.t. C1}: \sum_{j'=1}^{j} \mu_2^{j'} \leq \sum_{j'=1}^{j} \mu_1^{j'}, \forall j \in \{1, ..., L\},$$

$$\text{C2}: \mu_2^j \in \{0,1\}, \mu_1^j \in \{0,1\}, \forall j \in \{1, ..., L\},$$

$$\text{C3}: \sum_{j=1}^{L} \mu_1^j = 1, \sum_{j=1}^{L} \mu_2^j = 1.$$

$\mathcal{P}3$ is a standard mixed integer linear programming (MILP) problem. Since the number of CNN model layers is typically not very large, we can directly use an exhaustive search algorithm to calculate the minimum T_n and obtain μ_1^j and μ_2^j.

Finally, our proposed scheme, termed Layer Splitting and Computing Resource Allocation (LSCRA), conducts exhaustive search to ensure that all possible pairs of split layers are explored. Then, with each pair, we solve the optimal resource allocation based on bisection procedure from Eq. (18) to find the minimum delay attained. It is easy to see that LSCRA can obtain the optimal solution to $\mathcal{P}1$, and the computational complexity is $O(L^2 log F_s)$.

IV. Simulation Results

This section provides the numerical results to evaluate the learning performance of the proposed U-PSL framework and the effectiveness of the LSCRA algorithm and split layers strategy.

(a) HAM10000 under IID setting

(b) HAM10000 under non-IID setting

(c) MNIST under IID setting

(d) MNIST under non-IID setting

Fig. 3: Test accuracy of U-PSL, PSL, U-SFL, SFL on HAM10000 & MNIST dataset under IID/non-IID setting with $N = 5$, $F_s = 50$GHz.

A. Experiments Settings

In the simulations, we consider N clients randomly distributed around a wireless edge server. The computing capability of each client is uniformly distributed within $[0.5, 1.5]$ GHz, and the computing capability of the server is set to $[10, 50]$ GHz. The uplink data rate of each client is uniformly distributed within $[5, 30]$ Mbps, and the downlink data rate is set to $[2, 10] \times R_n^{\uparrow}$ Mbps. Other parameters can be found in Table III.

We evaluate the learning performance of the proposed U-PSL framework by deploying the ResNet-18 network on two image classification datasets, HAM10000 [21] and MNIST [22]. Furthermore, we conduct experiments under IID (independent and identically distributed) and non-IID data settings.

TABLE III: Parameter Settings

Parameter	value	Parameter	value
F_s	$[10, 50]$GHz	f_n	$[0.5, 1.5]$GHz
N	$[5, 100]$	β_n	64
K_s	$\frac{1}{32}$cycles/FLOPs	K_c	$\frac{1}{16}$cycles/FLOPs
$R_n^{\uparrow}$	$[5, 30]$Mbps	$R_n^{\downarrow}$	$[2, 10] \times R_n^{\uparrow}$Mbps

B. Performance Evaluation of the Proposed U-PSL Framework

In this subsection, we assess the performance of the proposed U-PSL framework in terms of test accuracy, convergence speed, training latency, and privacy preservation. We compare U-PSL with other distributed learning baselines, including PSL, SFL, and U-SFL, to examine the effectiveness of U-PSL. For fair comparison, the benchmark schemes also adopt optimal split layers and server computing resource allocation.

Fig. 4: Smashed data visualization.

Figure 3 demonstrates the test accuracy of these frameworks on the HAM10000 and MNIST datasets. It can be observed that U-PSL achieves a similar test accuracy compared to SFL, U-SFL and PSL as the models converge. Moreover, in some situations (e.g., Figure 3(a)), U-PSL requires the lowest time budget to reach a target accuracy. There are two reasons for this: One is that the client-side submodels in U-PSL are trained by user-specific data. Therefore, the client-side submodel may adapt better to user data in the early stages and perform better in terms of accuracy. The other is that U-PSL eliminates the need for model exchange between the clients and the server, reducing communication overhead and resulting in faster convergence compared to U-SFL and SFL.

Figure 4 illustrates the use of a raw image from HAM10000 to generate smashed data at the first and second cut layers, which are located after the skip connection of the third and fourth residual blocks in Resnet-18, respectively. From the visualization, the outputs significantly differ from the raw data. Also, it is hard to identify the label. In summary, *the U-PSL framework achieves both data and label privacy protection while achieving similar or even better performance compared to other benchmarks.*

C. Performance Evaluation of the Proposed LSCRA algorithm

In this subsection, we evaluate the performance of the proposed LSCRA scheme with respect to the server computing capacity and the number of service clients. We compare the proposed method with two benchmarks:

- **Benchmark a)**: *Optimal split layers & evenly allocated*, where the server and clients have the same cut layers as the proposed scheme, and the server computing resource is evenly allocated.
- **Benchmark b)**: *Suboptimal split layers & evenly allocated*, where the cut layers are set to the second performing case, and the server computing resource is evenly allocated.

Figure 5 illustrates the performance of the per-round training latency with respect to the server computing capacity. It can be observed that when the server's computing capacity is limited, the proposed scheme significantly reduces the training latency for each round. This is achieved by allocating more server computing resources to devices with weaker computing power and communication conditions.

Furthermore, when the server's computing capacity ranges from 10 GHz to 50 GHz, the proposed scheme ensures that the training time for each round does not decrease significantly. This is because, in scenarios where the server's computing capacity

Fig. 5: The performance for per-round training latency versus the server computing capacity with $[10, 50]$ GHz, $N = 100$.

Fig. 6: The performance for per-round training latency versus the number of clients from 10 to 100, $F_s = 50$ GHz

is sufficiently powerful, the communication time and the local training time of clients become the dominant factors. However, even in such cases, our method outperforms benchmark b), by finding the optimal split layers. This phenomenon demonstrates the importance of carefully selecting splitting layers and allocating computing resources. In a nutshell, our method reduces the training latency with varied computing capabilities, particularly in scenarios where the resources on the server are limited.

Figure 6 illustrates the performance of the per-round training latency with respect to the number of clients. As the number of clients increases, the time cost for each round associated with the two benchmarks shows a greater increase compared to our proposed scheme. This scenario aligns with real-world communication scenarios where a single server serves a large number of users.

V. Conclusions

In this paper, we proposed a novel split learning framework called U-Shaped Parallel Split Learning (U-PSL) to address model and label privacy preservation. By taking into account the additional communication overhead introduced by the U-shaped neural network, we have designed an effective resource allocation and layer splitting strategy to minimize the latency of U-PSL over wireless edge networks. Simulation results demonstrate that our proposed U-PSL framework retains a similar accuracy compared to existing SL benchmarks while preserving label privacy. Our results show the effectiveness and efficiency of adopting U-shaped SL at wireless edge networks. For the future work, we plan to derive the convergence results for our scheme and consider the joint optimization of computing resources and channel allocation for U-shaped PSL.

References

[1] X. Hou, J. Wang, Z. Fang, Y. Ren, K.-C. Chen, and L. Hanzo, "Edge intelligence for mission-critical 6G services in space-air-ground integrated networks," *IEEE Netw.*, vol. 36, no. 2, pp. 181–189, 2022.

[2] H. Peng and L.-C. Wang, "Energy Harvesting Reconfigurable Intelligent Surface for UAV Based on Robust Deep Reinforcement Learning," *IEEE Trans. Wireless Commun.*, 2023.

[3] X. Hou, J. Wang, Z. Fang, X. Zhang, S. Song, X. Zhang, and Y. Ren, "Machine-learning-aided Mission-critical Internet of Underwater Things," *IEEE Netw.*, vol. 35, no. 4, pp. 160–166, 2021.

[4] H. Peng, A.-H. Tsai, L.-C. Wang, and Z. Han, "LEOPARD: Parallel Optimal Deep Echo State Network Prediction Improves Service Coverage for UAV-Assisted Outdoor Hotspots," *IEEE Trans. Cogn. Commun. Netw.*, vol. 8, no. 1, pp. 282–295, 2021.

[5] J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated Learning: Strategies For Improving Communication Efficiency," *arXiv preprint arXiv:1610.05492*, 2016.

[6] X. Chen, G. Zhu, Y. Deng, and Y. Fang, "Federated Learning over Multihop Wireless Networks with In-Network Aggregation," *IEEE Trans. Wirel. Commun.*, vol. 21, no. 6, pp. 4622–4634, 2022.

[7] A. Imteaj, U. Thakker, S. Wang, J. Li, and M. H. Amini, "A Survey on Federated Learning for Resource-constrained IoT Devices," *IEEE Internet Things J.*, vol. 9, no. 1, pp. 1–24, 2021.

[8] P. Vepakomma, O. Gupta, T. Swedish, and R. Raskar, "Split Learning For Health: Distributed Deep Learning Without Sharing Raw Patient Data," *arXiv preprint arXiv:1812.00564*, 2018.

[9] O. Gupta and R. Raskar, "Distributed Learning of Deep Neural Network over Multiple Agents," *J. Netw. Comput. Appl.*, vol. 116, pp. 1–8, 2018.

[10] Z. Lin, G. Qu, X. Chen, and K. Huang, "Split Learning in 6G Edge Networks," *arXiv preprint arXiv:2306.12194*, 2023.

[11] Z. Lin, G. Zhu, Y. Deng, X. Chen, Y. Gao, K. Huang, and Y. Fang, "Efficient Parallel Split Learning over Resource-constrained Wireless Edge Networks," *arXiv preprint arXiv:2303.15991*, 2023.

[12] Z. Lin, G. Qu, Q. Chen, X. Chen, Z. Chen, and K. Huang, "Pushing Large Language Models to the 6G Edge: Vision, Challenges, and Opportunities," *arXiv preprint arXiv:2309.16739*, 2023.

[13] J. Jeon and J. Kim, "Privacy-sensitive Parallel Split Learning," in *Proc. ICOIN*, 2020.

[14] C. Thapa, P. C. M. Arachchige, S. Camtepe, and L. Sun, "Splitfed: When Federated Learning Meets Split Learning," in *Proc. AAAI*, 2022.

[15] B. Yin, Z. Chen, and M. Tao, "Predictive gan-powered multi-objective optimization for hybrid federated split learning," *IEEE Trans. Commun.*, 2023.

[16] Z. Yang, Y. Chen, H. Huangfu, M. Ran, H. Wang, X. Li, and Y. Zhang, "Robust Split Federated Learning for U-shaped Medical Image Networks," *arXiv preprint arXiv:2212.06378*, 2022.

[17] X. Chen, Y. Deng, H. Ding, G. Qu, H. Zhang, P. Li, and Y. Fang, "Vehicle as a service (VaaS): Leverage vehicles to build service networks and capabilities for smart cities," *arXiv preprint arXiv:2304.11397*, 2023.

[18] H. Ding and K. G. Shin, "Context-aware beam tracking for 5G mmwave V2I communications," *IEEE Trans. Mobile Comput.*, vol. 22, no. 6, pp. 3257 – 3269, June 2023.

[19] Z. Lin, L. Wang, J. Ding, B. Tan, and S. Jin, "Channel Power Gain Estimation for Terahertz Vehicle-to-infrastructure Networks," *IEEE Commun. Lett.*, vol. 27, no. 1, pp. 155–159, 2022.

[20] Z. Lin, L. Wang, J. Ding, Y. Xu, and B. Tan, "Tracking and Transmission Design in Terahertz V2I Networks," *IEEE Trans. Wireless Commun.*, 2022.

[21] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 Dataset, A Large Collection of Multi-source Dermatoscopic Images of Common Pigmented Skin Lesions," *Sci. Data*, vol. 5, no. 1, pp. 1–9, 2018.

[22] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based Learning Applied to Document Recognition," *Proc IEEE Inst Electr Electron Eng*, vol. 86, no. 11, pp. 2278–2324, 1998.