

Class-Wise Federated Averaging for Efficient Personalization

Gyujeong Lee
 SAKAK Inc.

regulation.lee@sakak.co.kr

Daeyeon Choi*
 The Cyber University of Korea

choidy@cuk.edu

Abstract

Federated learning (FL) enables collaborative model training across distributed clients without centralizing data. However, existing approaches such as Federated Averaging (FedAvg) often perform poorly with heterogeneous data distributions, failing to achieve personalization owing to their inability to capture class-specific information effectively. We propose Class-wise Federated Averaging (cwFedAvg), a novel personalized FL (PFL) framework that performs Federated Averaging for each class, to overcome the personalization limitations of FedAvg. cwFedAvg creates class-specific global models via weighted aggregation of local models using class distributions, and subsequently combines them to generate personalized local models. We further propose Weight Distribution Regularizer (WDR), which encourages deep networks to encode class-specific information efficiently by aligning empirical and approximated class distributions derived from output layer weights, to facilitate effective class-wise aggregation. Our experiments demonstrate the superior performance of cwFedAvg with WDR over existing PFL methods through efficient personalization while maintaining the communication cost of FedAvg and avoiding additional local training and pairwise computations.

1. Introduction

Deep networks demand large-scale training data for superior performance [15]; however, data collection faces significant challenges due to prohibitive costs and privacy constraints [11]. This limitation necessitates developing communication-efficient and privacy-preserving approaches to effectively utilize distributed data from sources such as data silos and edge devices [16, 27]. Federated learning (FL) addresses this challenge by enabling client collaboration through model aggregation without collecting client data [11, 18]. A foundational approach, Federated Averaging (FedAvg) [18], creates a single global

model by aggregating local models weighted by client sample sizes. However, this approach performs poorly with non-independent and identically distributed (non-IID) data due to its lack of personalization capability [9, 16, 32]. This limitation stems from how deep networks encode information. Deep networks develop pathways, where a pathway represents a union of weights from input to output [13], and encode class-specific information in these pathways [1, 12, 13]. Pathways demonstrate distinct patterns across different classes based on the class proportion [19, 25, 28]. However, FedAvg’s aggregation weighting factor, which only considers client sample sizes, fails to reflect class-specific pathways, limiting its personalization capability.

This study addresses the personalization limitation of FedAvg, motivated by class-specific pathways. We change the aggregation weighting factor of FedAvg using the class distributions of clients and create multiple global models specialized for each class using an adapted weighting factor. Based on this, we propose Class-wise Federated Averaging (cwFedAvg), a class-wise extension of FedAvg that performs Federated Averaging separately for each class. cwFedAvg implements a two-step aggregation process: first, we create *class-specific global models* by aggregating local models weighted by their respective client and class sample proportions; then, we generate personalized local models by aggregating these class-specific global models weighted by the class distribution of each client.

We identify two requirements to further improve the effectiveness of cwFedAvg: model weights (pathways) must strongly correlate with empirical class distributions to capture class-specific information effectively, and privacy-sensitive class distributions must be securely shared with the server. We analyze how empirical class distribution affects pathways and propose Weight Distribution Regularizer (WDR) to address these requirements. WDR strengthens the correlation between empirical class distributions and model weights by minimizing the distance between the empirical and approximated class distributions derived from the ℓ_2 -norms of output layer weight vectors. Additionally, it enables privacy-preserving sharing of class distributions using the approximated distribution as a proxy. Finally, we

*Corresponding author

extend the proposed approach by applying cwFedAvg selectively to upper layers, thereby reducing memory requirements while maintaining high performance.

Our work focuses on learning individualized models, one of the main approaches in personalized FL (PFL) [23]. This approach achieves personalization by modifying the model aggregation process, enabling adaptation to the unique data distribution of each client without post-training steps—a crucial advantage for resource-constrained devices [30]. Although the proposed method is a straightforward extension of FedAvg , it effectively addresses key limitations of existing learning individualized model approaches. Specifically, current methods either incur substantial overhead, such as weighted model aggregation [17, 31] and client-pair collaboration [10], or rely on strong assumptions about client group structures [2, 6, 7, 20]. Our proposed cwFedAvg delivers efficient personalization without requiring additional model downloads, pairwise computations, or clustering assumptions.

Our contributions and benefits of our approach are summarized as follows:

Our Contributions.

- We propose cwFedAvg , a novel PFL framework that performs Federated Averaging per class to create personalized models through class-specific global models.
- We develop WDR to enable effective class-wise model aggregation and ensure secure class distribution estimation.
- Our extensive experiments on four datasets and various levels of data heterogeneity demonstrate cwFedAvg 's superior performance over existing PFL methods and provide insights into personalization through visualization of output layer weight distributions and pathways.

Benefits of Our Approach.

- **Efficient Personalization:** cwFedAvg with WDR achieves personalization while maintaining the communication efficiency of FedAvg without requiring pairwise model collaborations or additional local training, making it particularly suitable for resource-constrained devices.
- **Enhanced Privacy:** Our distribution estimation method with WDR enables privacy-preserving class distribution sharing, making this approach applicable beyond personalization to other FL contexts, such as client selection.

2. Related Work

Personalized Federated Learning. Among PFL approaches, our work aligns with recent methods that personalize models through modified aggregation techniques. FedFomo [31] encourages FL among relevant clients by utilizing an optimally weighted combination of models. FedAMP [10] uses attention-based techniques to promote stronger collaboration between clients with similar data distributions. However, these methods often require heavy

computation for weight learning or additional communication to download other clients' models. Clustering-based FL methods address this issue by performing FL within client clusters. CFL [20] employs hierarchical clustering using cosine similarity of clients' gradient updates as a post-processing step. IFCA [7] assigns clients to pre-determined clusters and performs model aggregation within each cluster. Regularization-based methods eliminate both extensive computation and clustering assumptions. FedNH [4] adds normalization layers to ensure the uniformity and semantics of class prototypes. FedUV [21] introduces weight and representation regularization to emulate IID settings. Like these approaches, cwFedAvg leverages relevant clients for personalization with weight regularization. However, it avoids complex client collaboration and does not require assuming clients can be grouped into discrete clusters.

Correlating Model Parameters and Class Distribution.

In centralized machine learning, several studies have observed relationships between gradients or weights of deep networks and empirical class distribution [1, 12]. Anand et al. [1] revealed a correlation between the number of class samples and the magnitude of gradients associated with that class. They proposed an algorithm to accelerate learning by exploiting this correlation. Kang et al. [12] observed the positive proportional relationship between class and weight distribution. In the realm of FL, utilizing class distribution is pivotal as it can be employed for client selection [26] and loss function modification on clients [24, 29]. However, owing to privacy concerns, directly transmitting class distribution information to the server is typically prohibited. Consequently, several proxy methods have been proposed to estimate class distribution from a deep network. Yang et al. [26] utilized the gradient magnitude as a proxy to estimate the class distributions of clients and employed it for client selection. Wang et al. [24] developed a monitoring scheme that estimates class distribution based on the work of [1]. Our distribution estimation is motivated by Anand et al. [1] but uses weights instead of gradients and incorporates WDR to strengthen the class-weight distribution correlation.

3. Problem Formulation and Motivation

This section presents the problem formulation and motivation based on theoretical and empirical analysis.

3.1. Problem Formulation

The objective of traditional FL can be summarized as follows.

$$\min_{\mathbf{w}} f_G(\mathbf{w}) = \min_{\mathbf{w}} \sum_{i=1}^M p_i F_i(\mathbf{w}), \quad (1)$$

where $f_G(\cdot)$ and $F_i(\cdot)$ denote the global objective and the local objective of client i , respectively. The global objective $f_G(\mathbf{w})$ is the weighted sum of M local objectives, with

M being the number of clients. The weight p_i for each client is defined as the ratio of the number of data samples n_i on that client to the total number of data samples $n = \sum_{i=1}^M n_i$ across all clients, thus $p_i = \frac{n_i}{n}$. The local objective $F_i(\cdot)$ for each client i can be defined as the expected loss over the data distribution \mathcal{D}_i specific to that client. We approximate this expected loss using the empirical risk calculated over the local training data \mathcal{D}_i^{tr} available to the client. This empirical risk minimization is expressed as $\mathbb{E}_{z \sim \mathcal{D}_i} [\mathcal{L}(\cdot; z)] \approx \frac{1}{n_i} \sum_{z \in \mathcal{D}_i^{tr}} \mathcal{L}(\cdot; z)$, where z represents the data under local distribution \mathcal{D}_i .

In PFL, the global objective can take a more flexible form. The goal is to optimize a set of personalized models, one for each client, rather than a single global model. This objective can be expressed as follows.

$$\min_{\mathbf{W}} f_P(\mathbf{W}) = \min_{\mathbf{w}_i, i \in [M]} f_P(\mathbf{w}_1, \dots, \mathbf{w}_M), \quad (2)$$

where $f_P(\mathbf{W})$ denotes the global objective for the PFL algorithm, and \mathbf{w}_i denote personalized models. The goal is to find the optimal set of personalized models \mathbf{W}^* that minimizes the global objective function $f_P(\mathbf{W})$.

3.2. Motivation

The extension from FedAvg to cwFedAvg is made possible by methods to quantify class-specific information encoded in deep networks. Our work is motivated by empirical and theoretical findings showing that deep networks encode class-specific information in their weights and gradients [1, 12, 13]. A deep network develops pathways, where a pathway represents a union of paths (weights) from input to output [13]. Critical pathways, consisting of large-magnitude weights, demonstrate distinct patterns across different classes [19, 25, 28].

A Motivational Example. We first analyze how pathways diverge across clients due to non-IID data distributions through an example. We construct a synthetic dataset with two centers in \mathbb{R}^3 space generated from a Gaussian distribution for a binary classification task. The dataset contains 6000 samples: 3400 for class $k = 0$ and 2600 for class $k = 1$. The data are partitioned across three clients with distinct distributions. Client 1 exhibits a high imbalance, with 2700 samples of class $k = 0$ and 300 samples of class $k = 1$. Client 2 presents an inverse distribution, with 200 samples of class $k = 0$ and 1800 samples of class $k = 1$. Client 3 maintains a balanced distribution, with 500 samples per class. We employ a single-layer neural network with four neurons and ReLU activation functions.

We visualize the network architecture, where green and red lines represent positive and negative weights, respectively, and line width indicates magnitude. We draw critical pathways created from the top 8 out of 20 weights and denote \mathbf{w}_i^L as the i -th local model and \mathbf{w}^G as a global model

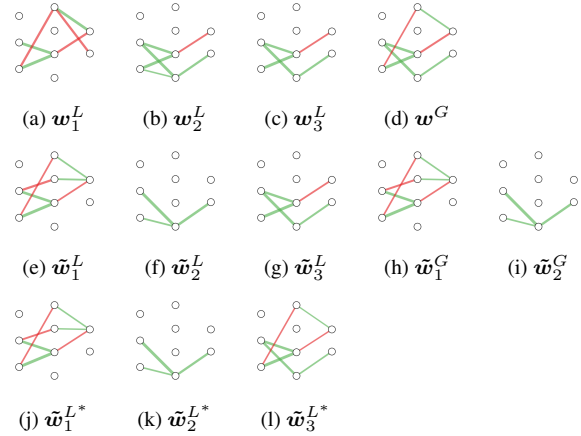


Figure 1. Pathway comparison of FedAvg and cwFedAvg. $\tilde{\mathbf{w}}$ denotes the model trained with WDR, and $*$ denotes the local models updated using class-specific global models (h) and (i).

of FedAvg. Figures 1a–1c show the learned pathways of local models for clients 1, 2, and 3, respectively. Notably, we observe that both output neurons contribute critical pathways in the three figures, and pathways in Figures 1b and 1c remain similar regardless of data imbalances. Figure 1d shows the pathways of the global model after the aggregation of FedAvg. This demonstrates why FedAvg fails to personalize, as the single averaged global model cannot capture the unique patterns of each client.

Theoretical Background and Observations. Although the magnitude of the weights in critical pathways varies with class proportions, its relationship with class distribution remains partially understood. The analysis of Anand et al. [1] revealed a correlation between gradients of weights in the output layer and class sample sizes. Let $\mathbf{w}_{i,j}$ denote the weight vector from penultimate layer neurons to output neuron j of client i . Anand et al. [1] formalized the correlation through a theorem: for a neural network classifier, the squared ℓ_2 -norms of the gradients of the weight vectors satisfy:

$$\frac{\mathbb{E} \|\nabla \mathcal{L}(\mathbf{w}_{i,j})\|_2^2}{\mathbb{E} \|\nabla \mathcal{L}(\mathbf{w}_{i,k})\|_2^2} \approx \frac{n_{i,j}^2}{n_{i,k}^2}, \quad (3)$$

where $n_{i,j}$ denotes the number of samples belonging to class j on client i .

Based on the theorem, we found a positive correlation between $\|\mathbf{w}_{i,j}\|_2$ and $n_{i,j}$. We further extend this relationship to demonstrate that the distribution derived from normalized ℓ_2 -norms of the weight vectors $\tilde{p}_{i,j}$ correlates with the empirical class distribution $p_{i,j} = \frac{n_{i,j}}{n_i}$. In this work, for K -class classification, $\tilde{p}_{i,j}$ is defined as:

$$\tilde{p}_{i,j} = \frac{\|\mathbf{w}_{i,j}\|_2}{\sum_{k=1}^K \|\mathbf{w}_{i,k}\|_2}. \quad (4)$$

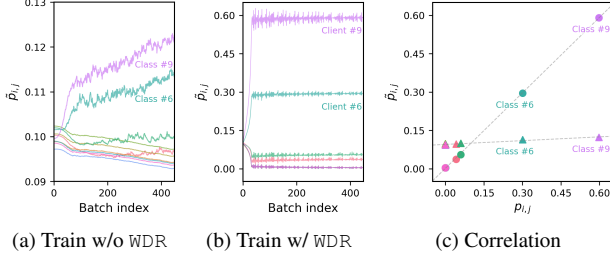


Figure 2. Evolution of $\tilde{p}_{i,j}$ and its correlation with $p_{i,j}$.

Indeed, we found that a correlation exists between them. We observed the evolution of $\tilde{p}_{i,j}$ for a client during the first FL round’s epoch on CIFAR-10, where each color represents a different class in Figure 2a. Classes 6 and 9 with larger samples exhibit higher $\tilde{p}_{i,j}$ values than the others. The triangular markers in Figure 2c indicate a positive correlation.

However, proving Eq. (3) and (4) for all network weights remains challenging. Unlike weights in the output layer that directly correspond to specific classes, weights in other layers lack this explicit class association. Nevertheless, we argue that these weights correlate with class proportions due to the cascading effect of backpropagation from the output layer. Based on this, we hypothesize that the contribution of each class to the trained model can be quantified as the product of weights and the class proportion $p_{i,j}w_i$. This is analogous to the quantification of the contribution of each client to the global model in FedAvg, quantified as p_iw_i .

4. The cwFedAvg Algorithm

In this section, we describe cwFedAvg with WDR and conclude by analyzing its relationship to FedAvg.

4.1. Class-Wise Aggregation

Class-Wise Local Model Aggregation. We denote w_i^L as the i -th local model and w_j^G as the j -th class-specific global model for M clients in a K -class classification task. The server aggregates local models by creating class-specific global models as follows (Figure 7c in the supplementary materials).

$$w_j^G = \sum_{i=1}^M q_{i,j} w_i^L \quad (5)$$

$$= \sum_{i=1}^M \frac{p_i \cdot p_{i,j}}{\sum_{i=1}^M p_i \cdot p_{i,j}} w_i^L, \quad (6)$$

where $q_{i,j}$ is a weighting factor. After expressing $q_{i,j}$ in terms of sample counts in Eq. (6), we get $q_{i,j} = \frac{n_{i,j}}{\sum_{i=1}^M n_{i,j}}$, which represents client i ’s proportion of total class j samples. Therefore, we consider this term as client i ’s contribution to class j in the system, and this aggregation becomes analogous to FedAvg performed separately for each class.

Class-Wise Global Model Aggregation. In contrast to FedAvg, where a single global model is copied to each local model, cwFedAvg performs a weighted summation of K class-specific global models as follows (Figure 7d in the supplementary materials).

$$w_i^L = \sum_{j=1}^K p_{i,j} w_j^G. \quad (7)$$

This weighted aggregation based on the class distributions of each client leads to personalization.

4.2. Weight Regularization for cwFedAvg

In this section, we argue that cwFedAvg requires two key conditions for effectiveness and privacy protection: (1) model weights that strongly correlate with the empirical class distribution to enable effective personalization, and (2) secure sharing of privacy-sensitive class distribution $p_{i,j}$ without compromising client information $n_{i,j}$. We present a method that simultaneously addresses both.

Regarding the first condition, we observed that $\tilde{p}_i = [\tilde{p}_{i,1}, \tilde{p}_{i,2}, \dots, \tilde{p}_{i,K}]$ correlates with the empirical distribution $p_i = [p_{i,1}, p_{i,2}, \dots, p_{i,K}]$, but they differ significantly as shown in Figure 2c (triangular markers). Notably, even when class j has zero or near-zero samples, $\tilde{p}_{i,j}$ maintains a non-negligible value approximately 0.1 (the leftmost triangles). To strengthen the correlation between p_i and \tilde{p}_i , we propose WDR, which minimizes the Euclidean distance between them:

$$\mathcal{R}_i = \|p_i - \tilde{p}_i\|_2. \quad (8)$$

With the regularization term, the total cost function $\tilde{\mathcal{L}}_i$ is denoted as $\tilde{\mathcal{L}}_i = \mathcal{L}_i + \lambda \mathcal{R}_i$, where $\lambda \in [0, \infty)$ is the regularization coefficient. When training with WDR (Figure 2b) $\tilde{p}_{i,j}$ evolves distinctly from the patterns in Figure 2a, with more apparent separation between classes. This improved separation is further illustrated in Figure 2c (circular markers), where WDR enables $\tilde{p}_{i,j}$ to accurately approximate $p_{i,j}$. This precise approximation enables replacing $p_{i,j}$ with $\tilde{p}_{i,j}$ as a reliable substitute, addressing our second condition. Clients, therefore, do not need to directly share $n_{i,j}$.

The pathway analysis provides additional validation of applying cwFedAvg with WDR. Figures 1e–1g show how WDR aligns weight patterns with the dominant class of each client, unlike Figures 1a–1c. For example, Figure 1e shows pathways connecting to the first output neuron with most samples from class 0, whereas Figure 1a shows pathways connecting to both output neurons. Through the class-wise aggregations, class-specific global models (Figures 1h and 1i) mirror the pathways of clients with the majority of samples for each respective class (Figures 1e and 1f). The updated client models (Figures 1j–1l) maintain their original pathway structures, showing minimal deviation from their pre-update models, thus indicating personalization.

Algorithm 1 cwFedAvg with WDR**Input:** M clients, regularization coefficient λ **Server executes:**

- 1: Initialize local model w_i^L and global model w_j^G
- 2: Initialize $\tilde{p}_{i,j}$ to $\frac{1}{K}$
- 3: **for** iteration $t = 1, \dots, T$ **do**
- 4: Sample a client subset \mathcal{C}^t
- 5: **for** client $i \in \mathcal{C}^t$ in parallel **do**
- 6: Compute w_i^L via Eq. (7) replacing $p_{i,j}$ with $\tilde{p}_{i,j}$
- 7: $w_i^L \leftarrow \text{ClientUpdate}(i, w_i^L)$
- 8: Compute $\tilde{p}_{i,j}$ via Eq. (4)
- 9: Compute w_j^G via Eq. (6) replacing $p_{i,j}$ with $\tilde{p}_{i,j}$

ClientUpdate(i, w_i^L):

- 1: **for** each local epoch **do**
- 2: **for** batch $b \in \mathcal{B}_i$ **do**
- 3: Update w_i^L according to the loss via Eq. (8)
- 4: **return** w_i^L to the server

Algorithm 1 presents the complete FL process of cwFedAvg with WDR. Building upon the FedAvg framework, the proposed method introduces two key features: class-wise aggregation (lines 6 and 9 on server side) and class distribution estimation using WDR (line 8 on server side and line 3 on client side).

4.3. Comparative Analysis with FedAvg

Insight on Relation between Global Models. We analyze class-specific global models of cwFedAvg compared with the global model of FedAvg under two scenarios. (1) In the IID case, one can prove that $w_j^G = w^G$, where w^G is the global model of FedAvg if all clients have the complete set of classes in their local datasets and maintain uniform class distribution. This implies that cwFedAvg performs similarly to FedAvg as data heterogeneity decreases. (2) In the extreme non-IID case, when each client i has data from exactly one class k_i , we demonstrate that

$$w_j^G = \sum_{i:k_i=j} \frac{n_i}{\sum_{l:k_l=j} n_l} w_i^L. \quad (9)$$

This equation indicates that the j -th class-specific global model is equivalent to the global model created by FedAvg of local models containing only class j . Both cases can be trivially proven through straightforward mathematical derivations. Although most practical scenarios fall between these two cases, theoretically proving how the differences between global models impact performance remains challenging.

Personalization by cwFedAvg. We visualize heatmaps of both the empirical data distribution and the ℓ_2 -norms of

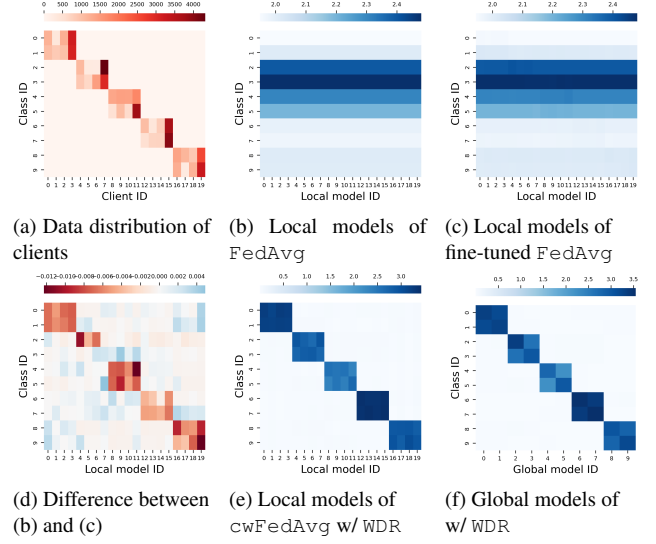


Figure 3. Heatmaps depicting the data distribution and ℓ_2 -norms of output layer weight vectors for the CIFAR-10 pathological setting. (a) Each cell represents the number of data samples belonging to class j for client i . (b)–(f) Each cell shows the ℓ_2 -norm of output layer weight vector, $\|\mathbf{w}_{i,j}\|_2$.

output layer weight vectors ($\|\mathbf{w}_{i,j}\|_2$) to demonstrate how cwFedAvg achieves effective personalization. Figure 3a presents the class-wise sample distribution across 20 clients for the CIFAR-10 pathological setting, where each client only has data from two classes. Next, Figures 3b (FedAvg) and 3c (fine-tuned FedAvg with local dataset) do not show any distinctions between local models in contrast to Figure 3a and are visually similar. However, a closer inspection reveals subtle differences between the two at positions corresponding to the diagonal darker areas in Figure 3a. Figure 3d visualizes the difference between the two heatmaps to highlight this distinction, where the diagonal areas appear darker, indicating that personalized models with local data (fine-tuned FedAvg) can induce changes in the norms. In contrast, Figure 3e (cwFedAvg with WDR) exhibits a pattern similar to Figure 3a, suggesting that each model has undergone personalization tailored to its possessed classes. As designed, each global model in Figure 3f specializes in specific classes, as each client possesses data from only two classes in this setting. Consequently, the norm can be a quantitative measure for assessing model personalization. Additional visualizations for more practical settings are provided in the supplementary materials.

Communication and Resource Requirements.

cwFedAvg maintains identical communication overhead to FedAvg, as all class-wise aggregations are performed on the server-side. This can be verified by comparing the dotted arrows in Figures 7a and 7c for local model aggregation, and Figures 7b and 7d for global model aggregation in the supplementary materials. The

Algorithm	Pathological setting		Practical setting ($\alpha = 0.1$)				
	CIFAR-10	CIFAR-100	MNIST	CIFAR-10	CIFAR-100	Tiny ImageNet	Tiny ImageNet*
FedAvg	60.68±0.84	28.22±0.32	98.70±0.04	61.94±0.56	32.44±0.42	21.35±0.12	24.71±0.15
FedProx	60.65±0.92	28.59±0.28	98.68±0.09	62.48±0.86	32.26±0.26	20.65±0.12	24.06±0.16
FedAMP	88.82±0.15	63.29±0.49	99.26±0.01	89.46±0.11	47.65±0.62	29.95±0.10	31.38±0.18
FedFomo	90.76±0.59	63.12±0.59	99.13±0.04	88.05±0.08	44.62±0.37	26.22±0.25	26.12±0.31
CFL	60.58±0.15	28.55±0.30	98.70±0.01	61.40±0.51	44.19±0.69	29.62±0.43	33.47±0.68
IFCA	72.84±4.80	58.98±2.38	99.10±0.06	70.12±0.13	34.86±1.02	19.93±0.59	26.68±0.16
FedNH	50.82±0.33	26.26±0.36	98.85±0.29	56.38±0.17	32.98±0.88	17.04±0.07	24.24±0.76
FedUV	88.11±0.13	62.72±0.28	99.25±0.09	88.59±0.09	46.80±0.20	28.09±0.06	25.45±0.03
cwFedAvg (Output)	91.23±0.04	67.50±0.14	99.52±0.03	88.65±0.19	56.29±0.18	41.38±0.12	43.51±0.14

Table 1. Classification accuracy (%) across datasets. Tiny ImageNet* indicates experiments using ResNet-18. cwFedAvg (Output) denotes cwFedAvg selectively applied to the output layer.

client-side operations in cwFedAvg preserve the same memory and computational requirements as FedAvg. When coupled with WDR, the only additional cost is the negligible computation for regularization. On the server side, although cwFedAvg requires more extensive storage to maintain multiple global models, its runtime memory allocation during aggregation matches FedAvg (Figure 7a and 7c). Memory efficiency can be achieved by applying cwFedAvg selectively to upper layers, and computational efficiency can be achieved through parallel execution of class-wise model aggregations.

Selective Application of cwFedAvg to Upper Layers. Deep networks exhibit layer-wise characteristics, where upper layers tend to learn class-specific features while lower layers capture more general, class-agnostic features. Based on this property, we examine a selective application of cwFedAvg. This approach applies cwFedAvg exclusively to the upper layers while maintaining FedAvg for the lower layers. We analyze the impact of this selective application on performance in the server, with detailed results presented in Section 5.

5. Experiments

In this section, we comprehensively evaluate cwFedAvg with WDR across various settings and analyze its convergence behavior against FedAvg.

5.1. Experimental Setup

Datasets. We evaluate our approach on four standard benchmark datasets: MNIST [5], CIFAR-10/100 [14], and Tiny ImageNet [3]. Each is partitioned into 75% training and 25% test splits. We examine two data heterogeneity settings: (1) pathological setting, where each client holds data from only a subset of classes—specifically two classes for MNIST and CIFAR-10, 10 classes for CIFAR-100, and 20 classes for Tiny ImageNet—and (2) practical setting, where client data distributions follow a Dirichlet distribution parameterized by α , with smaller values of α corresponding

to higher statistical heterogeneity across clients. We use $\alpha = 0.1$ as the default setting. Data distributions varying α are provided in the supplementary materials.

Evaluation Protocol. We measure performance using average test accuracy across all clients per FL round and report results from the best-performing round [10, 22, 30]. We compare our method against eight baselines across four categories: (1) traditional methods: FedAvg and FedProx, (2) aggregation-based methods: FedAMP and FedFomo, (3) clustering-based methods: CFL and IFCA, and (4) regularization-based methods: FedNH and FedUV.

Implementation. Our experiments employ two model architectures: a 4-layer CNN with ReLU activation functions (for MNIST, CIFAR-10/100, and Tiny ImageNet) and ResNet-18 [8] (for Tiny ImageNet). Following the settings of FedAvg [18], we set the training configuration to 20 clients as the default participating in each round, learning rate of 0.005, batch size of 10, and 1 local epoch. For regularizing clients with WDR, we set λ to 10 for MNIST and CIFAR-10, 1000 for CIFAR-100, and 2000 for Tiny ImageNet. Each experiment runs for 1,000 communication rounds to ensure convergence. All results are averaged over three independent runs with different random seeds. Experimental details are provided in the supplementary materials, and code is available at: <https://github.com/regulationLee/cwFedAvg>

5.2. Performance Comparison and Analysis

Classification Performance. Table 1 illustrates that PFL methods typically outperform traditional FL methods on non-IID data owing to the poor personalization ability of the global model. Among the PFL methods, cwFedAvg consistently outperforms all other settings except for one case in the CIFAR-10 practical setting, where the performance gap is minimal. Notably, cwFedAvg exhibits significant improvements over the others when the class count is large, typically indicating high data heterogeneity. The results for a text dataset are provided in the supplementary materials.

Algorithm	Comm. Cost	Number of clients		Data heterogeneity		
		50 Clients	100 Clients	$\alpha=0.01$	$\alpha=0.5$	$\alpha=1.0$
FedAvg	$2 \cdot \Sigma$	32.63 ± 0.34	32.32 ± 0.30	28.00 ± 0.92	36.18 ± 0.28	36.75 ± 0.34
FedProx	$2 \cdot \Sigma$	33.22 ± 0.20	32.64 ± 0.21	27.89 ± 0.24	35.93 ± 0.31	36.65 ± 0.39
FedAMP	$2 \cdot \Sigma$	44.97 ± 0.27	41.37 ± 0.35	73.46 ± 0.40	25.41 ± 0.14	21.23 ± 0.40
FedFomo	$(1+M) \cdot \Sigma$	42.62 ± 0.62	38.62 ± 0.08	71.30 ± 0.03	25.43 ± 0.58	18.95 ± 0.34
CFL	$2 \cdot \Sigma$	32.83 ± 0.78	32.88 ± 0.23	27.67 ± 0.17	38.32 ± 0.47	36.80 ± 0.07
IFCA	$(1+C) \cdot \Sigma$	29.17 ± 0.20	26.56 ± 0.45	53.89 ± 3.58	25.87 ± 0.57	22.27 ± 1.14
FedNH	$2 \cdot \Sigma$	33.14 ± 0.46	32.73 ± 0.24	25.48 ± 0.25	37.13 ± 0.41	20.41 ± 0.15
FedUV	$2 \cdot \Sigma$	44.30 ± 0.14	40.91 ± 0.22	72.67 ± 0.12	27.23 ± 0.25	37.41 ± 0.34
cwFedAvg (Output)	$2 \cdot \Sigma$	55.90 ± 0.35	53.54 ± 0.79	75.20 ± 0.21	40.78 ± 0.93	37.50 ± 0.10

Table 2. Communication cost formulation and classification accuracy (%) across different settings for CIFAR-100. Σ denotes total model parameters, and C denotes the number of clusters. cwFedAvg (Output) denotes cwFedAvg selectively applied to the output layer.

Applied algorithm and layer		Pathological setting		Practical setting ($\alpha = 0.1$)		
FedAvg	cwFedAvg	CIFAR-10	CIFAR-100	CIFAR-10	CIFAR-100	Tiny ImageNet
-	Conv1-Conv2-FC-Output	90.98 ± 0.15	65.91 ± 0.29	88.40 ± 0.13	54.99 ± 0.27	38.94 ± 0.38
Conv1	Conv2-FC-Output	90.99 ± 0.11	65.71 ± 0.19	88.55 ± 0.07	55.01 ± 0.25	38.76 ± 0.51
Conv1-Conv2	FC-Output	90.93 ± 0.06	65.22 ± 0.16	88.45 ± 0.06	54.98 ± 0.28	38.78 ± 0.68
Conv1-Conv2-FC	Output	91.23 ± 0.04	67.50 ± 0.14	88.65 ± 0.19	56.29 ± 0.18	41.38 ± 0.12
Conv1-Conv2-FC-Output	-	60.68 ± 0.84	28.22 ± 0.32	61.94 ± 0.56	32.44 ± 0.42	21.35 ± 0.12

Table 3. Classification accuracy (%) with selective application of cwFedAvg to a 4-layer CNN (Input-Conv1-Conv2-FC-Output).

Communication Cost. Table 2 presents the communication cost formulation per iteration. cwFedAvg maintains the same communication overhead as FedAvg as explained in Section 4. In contrast, other PFL approaches often incur larger communication costs: FedFomo and IFCA require downloading additional models, which increases downstream communication.

Client Scalability. Table 2 demonstrates the performance scaling to the number of clients. All PFL methods exhibit performance degradation as the number of clients increases. However, cwFedAvg maintains high performance even with larger client counts compared with other methods, thus beneficial for large-scale deployments of devices.

Data Heterogeneity and Convergence Analysis. We evaluate the model accuracy under non-IID data distributions controlled by the Dirichlet distribution parameter α . Table 2 shows that cwFedAvg maintains robust performance across different α values. In contrast, several PFL methods underperform compared with traditional approaches as data becomes more IID (larger α). Moreover, the performance of cwFedAvg converges to that of FedAvg as α increases, which validates our analysis of global model relationships in Section 4.3. Figure 4 shows the average training loss for 20 clients under varying data heterogeneity. The results demonstrate that cwFedAvg converges significantly faster than FedAvg on both CIFAR-10 and CIFAR-100 datasets when data are highly heterogeneous ($\alpha = 0.01$ and 0.1).

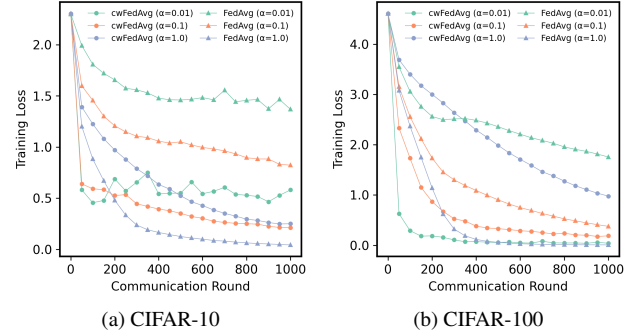


Figure 4. Convergence comparison of cwFedAvg and FedAvg.

This aligns with our pathway analysis in Figure 1, as local models in cwFedAvg adapt to their class distributions. However, when data is nearly IID ($\alpha = 1.0$), FedAvg converges faster. We provide per-client training loss comparisons in the supplementary materials.

Selective Application of cwFedAvg. Unlike FedAvg, PFL methods often require additional memory for storing multiple models, such as cluster models in IFCA and class-specific global models in cwFedAvg. We selectively apply cwFedAvg to upper layers only to reduce memory overhead. Across various datasets, applying cwFedAvg solely to the output layer achieves optimal performance (Table 3). This result aligns with the conventional understanding that upper layers capture class-specific features. Memory cost comparison is provided in the supplementary materials.

Distribution	WDR	CIFAR-100	Tiny ImageNet
Empirical (\mathbf{p})	✗	48.27±0.74	31.09±0.14
Approximation ($\tilde{\mathbf{p}}$)	✗	32.25±0.56	20.08±0.96
Approximation ($\hat{\mathbf{p}}$)	✓	54.99±0.27	38.94±0.38

Table 4. Ablation study results showing classification accuracy (%) of cwFedAvg under practical settings.

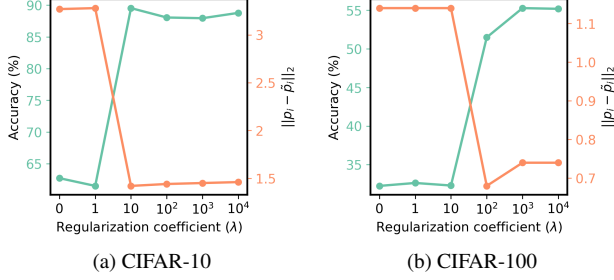


Figure 5. Influence of regularization coefficient (λ) on accuracy and regularization term ($\|\mathbf{p}_i - \tilde{\mathbf{p}}_i\|_2$) in practical settings.

Ablation Study. We examine the effect of empirical data distribution \mathbf{p} , approximated distribution $\tilde{\mathbf{p}}$, and WDR in practical settings when cwFedAvg is applied to all layers. First, as expected, with $\tilde{\mathbf{p}}$ and without WDR, cwFedAvg (the second row of Table 4) shows performance comparable to FedAvg (the first row of Table 1) as it fails to encode and extract class-specific information properly. Notably, cwFedAvg with $\tilde{\mathbf{p}}$ and WDR significantly outperforms cwFedAvg with \mathbf{p} without WDR when the class count is large. This result suggests that WDR effectively approximates \mathbf{p} while regularizing pathways to be class-specific.

Impact of Regularization Coefficient. Figure 5 illustrates the impact of regularization coefficient (λ) on accuracy and the regularization term ($\|\mathbf{p}_i - \tilde{\mathbf{p}}_i\|_2$) for CIFAR-10/100 practical heterogeneous settings. Increasing λ reduces the regularization term, enabling cwFedAvg to effectively utilize $\tilde{\mathbf{p}}_i$ for class-wise aggregation. Based on these results, we select the optimal λ that maximizes regularization strength while maintaining accuracy.

5.3. Personalization for Many-Class and Highly Imbalanced Data

Figure 6 shows heatmaps of the empirical data distribution and the ℓ_2 -norms of output layer weight vectors ($\|\mathbf{w}_{i,j}\|_2$) for the CIFAR-100 practical setting. The results demonstrate similar patterns observed in the CIFAR-10 pathological setting (Figure 3). Notably, Figure 6c (FedAMP) exhibits a pattern similar to Figure 6e (cwFedAvg with WDR), which aligns with their shared personalized aggregation-based approach, without FedAMP’s explicit regularization of output layer weights. Figure 6d (IFCA), a clustering-based method, reveals distinct cluster formations but does

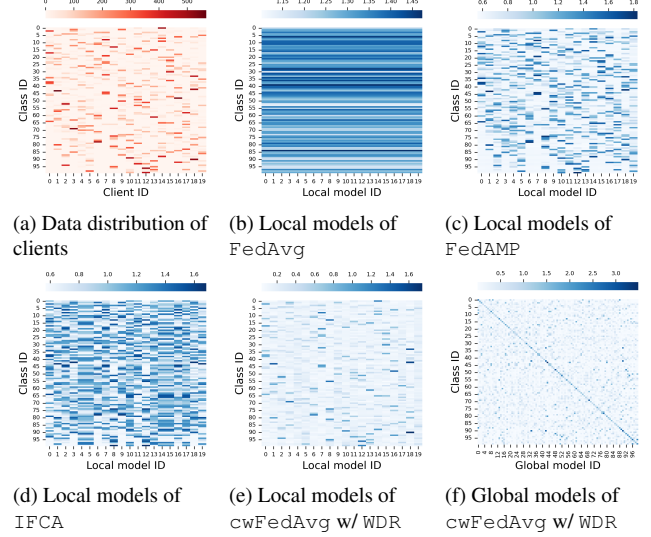


Figure 6. Heatmaps depicting the data distribution and ℓ_2 -norms of output layer weight vectors for the CIFAR-100 practical setting. (a) Each cell represents the number of data samples belonging to class j for client i . (b)–(f) Each cell shows the ℓ_2 -norm of output layer weight vector, $\|\mathbf{w}_{i,j}\|_2$.

not show personalization patterns. Figures 6e and 6f confirm that our class distribution estimation with WDR works appropriately even for many-class and highly imbalanced data settings. Additional visualizations for other settings are provided in the supplementary materials.

6. Limitation and Conclusion

The cwFedAvg method requires storing multiple global models on the server, equal to the number of classes. This requires more server memory than FedAvg, but the overhead can be significantly reduced by applying cwFedAvg to only upper layers with higher performance.

Despite its simplicity, cwFedAvg achieves efficient personalization and provides significant advantages over existing algorithms in cross-device PFL scenarios. cwFedAvg eliminates the overhead of additional client-side training or downloading other clients’ models that conventional PFL methods often require. The computational complexity of the algorithm scales linearly with the number of clients, whereas other PFL methods often require quadratic complexity owing to pairwise information exchange. This scalability makes cwFedAvg particularly suitable for large-scale deployments, as real-world applications typically involve a fixed number of classes while the number of participating clients grows. Furthermore, as an aggregation module, cwFedAvg with WDR can be integrated into existing PFL frameworks as a replacement for FedAvg, offering potential performance enhancements.

Acknowledgments This research was supported by Seoul R&BD Program (VC240022) through the Seoul Business Agency (SBA) funded by Seoul Metropolitan Government and supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2023-00277078).

References

- [1] Rangachari Anand, Kishan G Mehrotra, Chilukuri K Mohan, and Sanjay Ranka. An improved algorithm for neural network classification of imbalanced training sets. *IEEE transactions on neural networks*, 4(6):962–969, 1993. 1, 2, 3
- [2] Christopher Briggs, Zhong Fan, and Peter Andras. Federated learning with hierarchical clustering of local updates to improve training on non-iid data. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE, 2020. 2
- [3] Patryk Chrabaszcz, Ilya Loshchilov, and Frank Hutter. A downsampled variant of imagenet as an alternative to the cifar datasets. *arXiv preprint arXiv:1707.08819*, 2017. 6
- [4] Yutong Dai, Zeyuan Chen, Junnan Li, Shelby Heinecke, Lichao Sun, and Ran Xu. Tackling data heterogeneity in federated learning with class prototypes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7314–7322, 2023. 2
- [5] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012. 6
- [6] Moming Duan, Duo Liu, Xinyuan Ji, Renping Liu, Liang Liang, Xianzhang Chen, and Yujuan Tan. Fedgroup: Efficient federated learning via decomposed similarity-based clustering. In *2021 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)*, pages 228–237. IEEE, 2021. 2
- [7] Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. An efficient framework for clustered federated learning. *Advances in Neural Information Processing Systems*, 33:19586–19597, 2020. 2
- [8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6
- [9] Kevin Hsieh, Amar Phanishayee, Onur Mutlu, and Phillip Gibbons. The non-iid data quagmire of decentralized machine learning. In *International Conference on Machine Learning*, pages 4387–4398. PMLR, 2020. 1
- [10] Yutao Huang, Lingyang Chu, Zirui Zhou, Lanjun Wang, Jiangchuan Liu, Jian Pei, and Yong Zhang. Personalized cross-silo federated learning on non-iid data. In *Proceedings of the AAAI conference on artificial intelligence*, pages 7865–7873, 2021. 2, 6
- [11] Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2): 1–210, 2021. 1
- [12] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations (ICLR)*, 2020. 1, 2, 3
- [13] Ashkan Khakzar, Soroosh Baselizadeh, Saurabh Khanduja, Christian Rupprecht, Seong Tae Kim, and Nassir Navab. Neural response interpretation through the lens of critical pathways. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13528–13538, 2021. 1, 3
- [14] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6
- [15] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015. 1
- [16] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine learning and systems*, 2:429–450, 2020. 1
- [17] Jun Luo and Shandong Wu. Adapt to adaptation: Learning personalization for cross-silo federated learning. In *IJCAI: proceedings of the conference*, page 2166. NIH Public Access, 2022. 2
- [18] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017. 1, 6
- [19] Yuxian Qiu, Jingwen Leng, Cong Guo, Quan Chen, Chao Li, Minyi Guo, and Yuhao Zhu. Adversarial defense through network profiling based path extraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4777–4786, 2019. 1, 3
- [20] Felix Sattler, Klaus-Robert Müller, and Wojciech Samek. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE transactions on neural networks and learning systems*, 32(8):3710–3722, 2020. 2
- [21] Ha Min Son, Moon-Hyun Kim, Tai-Myoung Chung, Chao Huang, and Xin Liu. Feduv: Uniformity and variance for heterogeneous federated learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5863–5872, 2024. 2
- [22] Canh T Dinh, Nguyen Tran, and Josh Nguyen. Personalized federated learning with moreau envelopes. *Advances in Neural Information Processing Systems*, 33:21394–21405, 2020. 6
- [23] Alysa Ziyang Tan, Han Yu, Lizhen Cui, and Qiang Yang. Towards personalized federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 2
- [24] Lixu Wang, Shichao Xu, Xiao Wang, and Qi Zhu. Addressing class imbalance in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10165–10173, 2021. 2

- [25] Yulong Wang, Hang Su, Bo Zhang, and Xiaolin Hu. Interpret neural networks by identifying critical data routing paths. In *proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8906–8914, 2018. [1](#), [3](#)
- [26] Miao Yang, Ximin Wang, Hongbin Zhu, Haifeng Wang, and Hua Qian. Federated learning with class imbalance reduction. In *2021 29th European Signal Processing Conference (EUSIPCO)*, pages 2174–2178. IEEE, 2021. [2](#)
- [27] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–19, 2019. [1](#)
- [28] Fuxun Yu, Zhuwei Qin, and Xiang Chen. Distilling critical paths in convolutional neural networks. *arXiv preprint arXiv:1811.02643*, 2018. [1](#), [3](#)
- [29] Jie Zhang, Zhiqi Li, Bo Li, Jianghe Xu, Shuang Wu, Shouhong Ding, and Chao Wu. Federated learning with label distribution skew via logits calibration. In *International Conference on Machine Learning*, pages 26311–26329. PMLR, 2022. [2](#)
- [30] Jianqing Zhang, Yang Hua, Hao Wang, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. Fedala: Adaptive local aggregation for personalized federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 11237–11244, 2023. [2](#), [6](#), [1](#)
- [31] Michael Zhang, Karan Sapra, Sanja Fidler, Serena Yeung, and Jose M Alvarez. Personalized federated learning with first order model optimization. *arXiv preprint arXiv:2012.08565*, 2020. [2](#)
- [32] Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018. [1](#)