Logit Calibration for Non-IID and Long-Tailed Data in Federated Learning

1st Huan Wang
School of Cyber Engineering
Xidian University
Xi'an, China
wh_1018@stu.xidian.edu.cn

2nd, * Lijuan Wang
School of Cyber Engineering
Xidian University
Xi'an, China
ljwang@xidian.edu.cn

3rd Jun Shen
School of Computing & Information Technology
University of Wollongong
Wollongong, Australia
jshen@uow.edu.au

***Abstract*—Federated learning (FL) strives to enable collaborative training of deep models on the distributed clients of different data without centrally aggregating raw data and hence improving data privacy. Nevertheless, a central challenge in training classification models in the federated system is learning with non-IID data. Most of the existing work is dedicated to eliminating the heterogeneous influence of non-IID data in a federated system. However, in many real-world FL applications, the co-occurrence of data heterogeneity and long-tailed distribution is unavoidable. The universal class distribution is long-tailed, causing them to become easily biased towards head classes, which severely harms the global model performance. In this work, we also discovered an intriguing fact that the classifier logit vector (*i.e.*, pre-softmax output) introduces a heterogeneity drift during the learning process of local training and global optimization, which harms the convergence as well as model performance. Therefore, motivated by the above finding, we propose a novel logit calibration FL method to solve the joint problem of non-IID and long-tailed data in federated learning, called Federated Learning with Logit Calibration (FedLC). First, we presented a method to mitigate the local update drift by calculating the Wasserstein distance among adjacent client logits and then aggregating similar clients to regulate local training. Second, based on the model ensemble, a new distillation method with logit calibration and class weighting was proposed by exploiting the diversity of local models trained on heterogeneous data, which effectively alleviates the global drift problem under long-tailed distribution. Finally, we evaluated FedLC using a highly non-IID and long-tailed experimental setting, comprehensive experiments on several benchmark datasets demonstrated that FedLC achieved superior performance compared with state-of-the-art FL methods, which fully illustrated the effectiveness of the logit calibration strategy.**

***Index Terms*—long-tailed, non-IID, drift, logit calibration**

I. Introduction

Machine learning has been widely used in many applications in people's daily life, in order to obtain satisfactory performance, usually enough client data is required for model training [1]. Unfortunately, with the increasing awareness of privacy and security, only limited or poor quality data from different organizations is available [2]. In this situation, federated learning (FL) [3], [4] allows multiple local clients to collaboratively train a globally shared model with data privacy well-protected. In FL, a service provider (the server) can communicate with distributed data sources (the local clients), while the clients hold private data separately. The global model on the server is produced by aggregating local models transmitted from clients without the requirement of transferring any data from them. However, a high degree of systematic and statistical heterogeneity remains a major challenge for federated learning versus traditional distributed optimization [5]. Since the differences among different clients, local data samples are often not independent and identically distributed (non-IID), training on this kind of data results in poor performance and generalization ability of the FL global model. In addition, the distribution of each local client dataset is highly different from the global distribution, which leads to the local optimization objective of each client is inconsistent with the global optima [6], resulting in the non-IID heterogeneous drift problem.

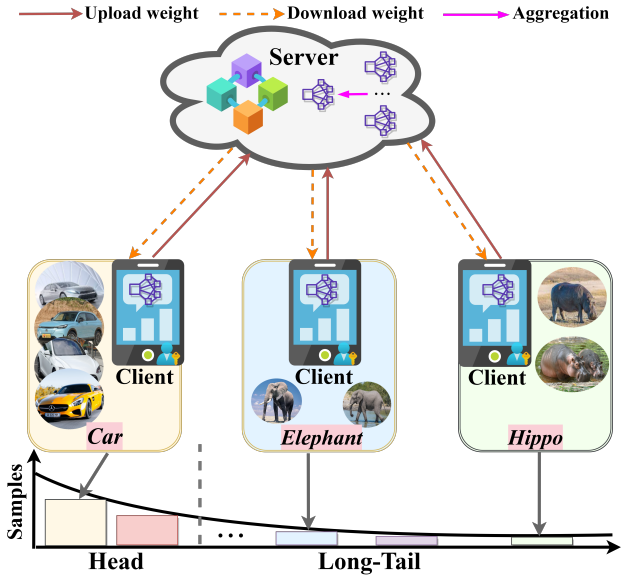


Fig. 1. An example of federated learning framework with long-tailed data. A few classes like car have a large number of samples, while many classes like elephant or hippo only take a small portion.

Previous literature has both empirically and theoretically demonstrated that heterogeneous drift makes the model damaged and unstable [7], [8]. Due to the specificity of data among local clients, in the local training stage, the local models are updated towards the local optima, which may be far from the

* To whom correspondence should be addressed.

global optimum. Furthermore, the averaged global model can also be far from a convergent optimal solution [6]. In the recent studies, a number of methods have been proposed to deal with non-IID data and drift problem in federated learning. They can be roughly categorized into client-side methods and server-side methods, mainly from two complementary perspectives: the former aims to improve the local training process, by adjusting the local model to limit the discrepancy between the client model and the global model, thus stabilizing local training [9], [10]. The latter adopts to ameliorate the model aggregation mechanism, thereby alleviating the global drift issue induced by heterogeneity [11].

Although most existing FL methods have made great progress in addressing the issues of stability, client drift, and heterogeneous label distribution. However, in the real-world scenario, training data tends to be heterogeneous and long-tailed class distribution. In the setting of a long-tailed distribution, the training data is usually heavily class-imbalanced, where the number of samples in a few classes (called head classes) severely outnumbers that in some other classes (called tail classes). As shown in Fig. 1, the trained classification model tends to be biased towards head classes with a large amount of training data (such as car class), resulting in poor performance of the model on tail classes with limited data (such as elephant or hippo class). This class imbalance in the number of training samples makes the training of recognition models based on deep networks very challenging. How to effectively model data with long-tailed class distribution is called long-tail learning [12], [13], which has been extensively studied in recent years. Existing methods can be mainly divided into one-stage and two-stage imbalance learning. The former uses the idea of class-balanced resampling to balance the contribution of each class [14], while the latter solves the problem of data imbalance based on the method of decoupling representation learning and classifier retraining [15].

Local or global class imbalance is the norm rather than the exception in many FL application tasks, if the training data across clients is long-tailed and heterogeneous at the same time, the joint problem becomes complicated and challenging. Some literature proposes introducing traditional imbalanced learning methods on the client to alleviate the influence of class distribution imbalance [16], [17]. However, merely regulating local training raises a drift across client updates, and this approach can not fully leverage the diverse knowledge across user models. Meanwhile, local heterogeneity is not adequately addressed, which may in turn affect the quality of the aggregated global model. There are also methods specifically designed for FL with long-tailed data [18]. Unfortunately, long-tailed data may be unbalanced at both the local and global levels, and these methods often only consider parts.

In this paper, we find that the heterogeneity and long-tailed distribution across clients lead to local drift in the classifier logit, which further induces the global model update bias. Motivated by the idea of model retraining and the drift concerns of FL, we consider both local and global updates and propose a novel client-server FL method called Federated Learning with Logit Calibration (FedLC) to effectively solve the joint problem of data heterogeneity and long-tailed class distribution. First, to mitigate local drift at the local update, we propose local logit calibration, which regulates local training by computing the Wasserstein distance among client logits to aggregate the similar clients. Then for the global drift, although weighted parameter aggregation is widely adopted, the setting of long-tailed distribution leads to poor global model performance. Taking advantage of the diversity of local models, we propose a distillation method with logit adjustment and class weighting based on model ensemble, which effectively alleviates the negative influence of global update drift. Specifically, we transfer knowledge from the client-side ensemble model to the global model via logit distillation. However, the long-tailed effect may still exist, so we calibrate the global model logit by adjusting the importance weights of the head and tail classes to eliminate class bias. Finally, we also perform distillation on the auxiliary dataset to further enhance the performance of the global model. The logit-calibrated global model mitigates the distance between local and global updates well, while generalizing well on both head and tail class data.

Our main contributions are highlighted as follows:

- We study this challenging FL problem based on heterogeneous and long-tailed data from the perspective of local and global update drift.
- We propose to aggregate similar clients by computing the Wasserstein distance to correct local logit, which effectively mitigate local update drift.
- We propose an ensemble distillation method with logit adjustment and class weighting, which effectively alleviates the negative influence of global update drift.
- We conduct extensive experimental studies to evaluate FedLC, and our FedLC achieves superior performance compared to state-of-the-art FL methods.

II. Related Work

Federated Learning: Statistical heterogeneity is common with data being non-identically distributed between devices. There have been many methods proposed trying to improve the problem of data heterogeneity in FL, which are mainly from two groups: one focuses on stabilizing local training from a client-side perspective, by adjusting the deviation of the local models and global model over the parameter space to address local drift [7]. Specifically, penalizing client-side updates through regularization to mitigate weight differences [9], using data pre-processing to reduce data heterogeneity [19], and learning affine transformations under the assumption that the data follows an affine distribution [20]. Another aims to improve the efficacy of model aggregation through deep learning methods from a server perspective. FedMA [21] proposed an aggregation strategy for non-IID data partition that shares global model in a layer-wise manner. Recently, knowledge distillation has emerged as an effective solution to alleviate the model drift problem by integrating knowledge for better local optimization and global aggregation [22]. In

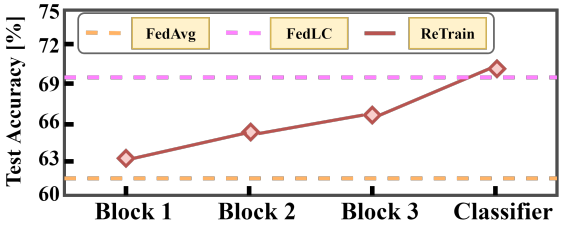
addition, some other works try to obtain a unique model for each client to combat the ill-effects of non-IID by retaining personalized information [23].

Long-Tail Learning: The long-tailed class imbalance is a common problem in real-world visual recognition tasks, where models are easily biased towards the dominant classes and underperform on the tail classes. Existing research on long-tail learning is mainly divided into three perspectives: re-sampling, re-weighting, and transfer learning methods [12], [13]. The re-sampling method mainly includes undersampling of the head class samples and oversampling of the tail class samples. BBN [24] learned from both original and augmented data through a dual-branch model, and [25] improved the long-tail recognition issue by decoupling representations and performing class-balanced sampling learning on the classifier. The re-weighting aims to solve the long-tail imbalance problem by improving the loss function. [26] proposed that with the increase of the samples, the income brought by each sample decreases significantly, and designed a better class-balanced loss. Further, [27] designed a re-weighted loss based on a two-step training method to achieve better long-tail classification effect. In addition, [28] and [29] retrain the classifier to mitigate the long-tail effect. However, we are different, eliminating class distribution imbalance and non-IID heterogeneous drift issues by correcting biased logits.

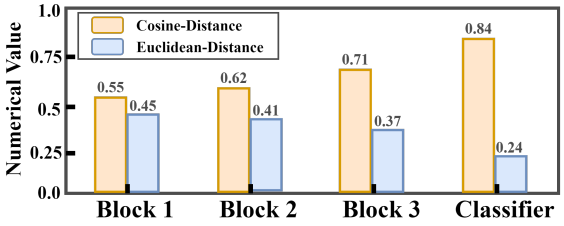
III. Our Approach: FedLC

To address the joint problem of heterogeneous and long-tailed data in FL, we propose an effective new federated learning approach of FedLC based on logit calibration. We first describe the problem setting with some basic notations, and then further emphasize our finding: the classifier logit severely harms model performance. Next, we describe the overall FedLC framework, introducing local logit calibration and global logit distillation to reduce local drift and global drift under long-tailed distribution, respectively.

Notations and Preliminaries. We consider a typical FL setting for solving the general problem of multi-class classification tasks. Given a dataset $\mathcal{D} = \{x_i, y_i\}_{i=1}^{N}$, where x_i is the training data instance in $\mathcal{X}$ and has the corresponding category label $y_i \in \{1, \ldots, \mathcal{C}\}$, N is the total number of samples and $\mathcal{C}$ is the number of classes. In the FL system, K clients holding privately non-IID datasets $\mathcal{D}^1$,..,$\mathcal{D}^K$, respectively. Furthermore, frequency of data samples from $\mathcal{D}$ follow a long-tailed distribution, that a small portion of clients owning most of the data, as an example shown in Fig. 1. Note that the private non-IID dataset $\mathcal{D}^k$ on the local client k also conforms to the long-tailed distribution. For the baseline model θ_w with parameter w in FL, we treat the last MLP layer as the classifier l_w (outputs a class confidence logit vector) and all previous layers as the feature extractor f_w (outputs a feature representation with dimension d). We define the FL global model as s_w. For each local client k, its private dataset is $\mathcal{D}_k$, the local model is w_k, and the local client average logits is denoted as ϕ_k.



(a) Test accuracy for FedAvg and retrained blocks.



(b) Numerical values for cosine and euclidean distances.

Fig. 2. (a) The performance of each block retraining based on the pre-trained FedAvg model; (b) The cosine and euclidean distances between pre-trained FedAvg model average logits and retrained block average logits.

A. Motivation for FedLC

For classification tasks with heterogeneous and long-tailed data, the distribution of sample features and category annotations are inherently uncoupled, improving data samples based on re-sampling should only rebalance the classifier. We empirically indicate that for the joint problem of FL and long-tailed distribution, cross-client classifier updates introduce local heterogeneous drift, and the global model after further aggregation is also negatively affected by global drift. Meanwhile, the biased classifier on the client is more biased towards the dominant class, resulting in the poor performance of the FL global model. Especially, training the model with non-IID and long-tailed data leads to the logits overconfidence, and different layers of the model suffer from the biased logits.

We experimentally verify the view: cross-client classifier logits introduce the heterogeneous drift problem in the local training and global aggregation phases, while the biased logits focus more on the head classes, which causes the model to be seriously biased. First, we divide the ResNet-8 model into four parts: the feature extractor is split into three blocks, and the classifier is a separate block. Then sample from the training data (2% of the data size) for each class as an auxiliary dataset to retrain each block while freezing the others. The FedAvg model is pre-trained on CIFAR-10 dataset [30]. In the whole process, we set the IF (the ratio between the number of samples in the largest class and that in the smallest class) to 50 and the non-IID heterogeneity degree α (dirichlet distribution) to 0.4, and then evaluated on the test dataset. As shown in Fig. 2(a), we observe that retraining any block improves performance to a certain extent, as the auxiliary dataset corrects the bias in the model updates. Further, we also calculated the distances between the average logits after retraining each block and the average logits of the FedAvg model, as illustrated in Fig. 2(b). Note that retraining the

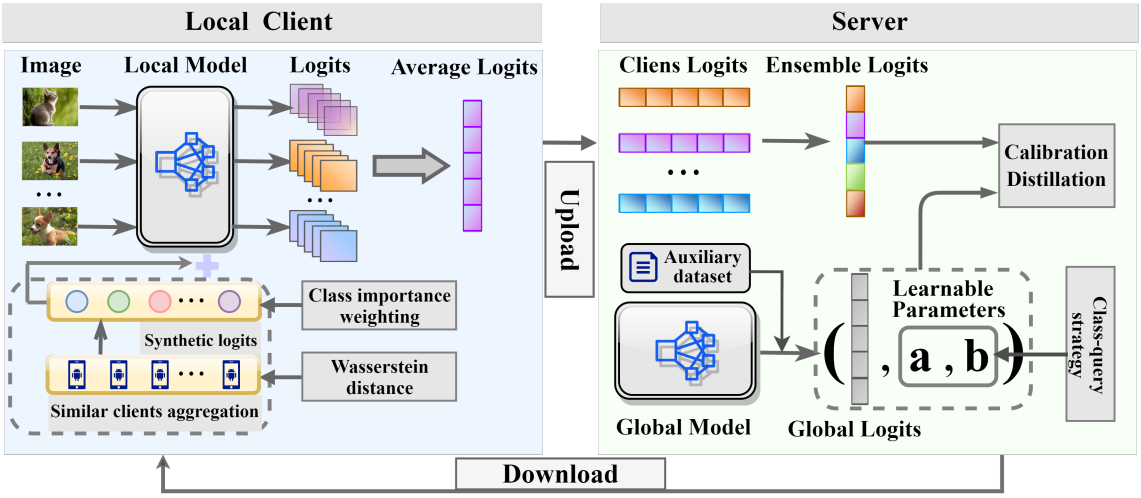


Fig. 3. The framework of FedLC.

classifier achieves the highest performance gain (+9.11%) and has the smallest distance from FedAvg model average logits.

Therefore, it can be seen that the cross-client classifier with biased logits is the main factor contributing to the poor performance of the FL global model. Motivated by the above finding, FedLC effectively alleviates the performance damage and drift issues caused by the classifier biased logits through logit calibration.

B. Logit Calibration Framework

Our core idea is to correct the biased logits during local training and global aggregation to alleviate the drift of local and global model updates, thereby addressing the heterogeneous and long-tailed distribution problem in FL.

FedLC is based on client-server architecture and proposes corresponding solutions to the heterogeneous long-tail problem through logit calibration in local and global perspectives, respectively. Specifically, the local client aggregates the logits of similar clients by calculating Wasserstein distance and then weights the average to obtain synthetic logits. The local model learns generalization knowledge from synthetic logits to mitigate local drift and eliminates the influence of local long-tailed distribution through a class-importance weighting strategy. Based on knowledge distillation in the global aggregation stage, the diverse knowledge of heterogeneous data is transferred from the client ensemble logits to the global model, thus eliminating global drift. To further alleviate the long-tail effect of the global model, we introduce learnable parameters to calibrate the global model logits through a class-query strategy. FedLC decouples heterogeneous long-tailed data problems from local and global perspectives and achieves excellent performance, which is an effective solution to obtaining a promising FL global model. The architecture of FedLC is provided in Fig. 3.

Local Logit Calibration. Local private data is heterogeneous and long-tailed, and we address local drift and long-tail effects through local logit calibration. Local logit calibration consists of two core components: learning generalization knowledge from synthetic logits to mitigate local drift caused by local model updates and eliminating local long-tailed effects through a class-importance weighting strategy. First, to obtain the synthetic logits, each client sends the average logits of local data to the server when the local update is completed. Note that since all logits of the local data are averaged, no relevant privacy information is leaked. For the client k, it receives the average logits of each client model from the server and obtains the distance between the local data distribution and other client data distributions by calculating Wasserstein distance. The Wasserstein distance represents the similarity among clients, and then the client k aggregates the average logits of adjacent clients to obtain the synthetic logits. Now we can calculate the similarity of logits between two clients i and j by the Wasserstein distance, but to avoid expensive computation, we approximate the Wasserstein distance d as:

$$d_{i,j} = \sum_{c=1}^{\mathcal{C}} \left(\left\| \mu_i^c - \mu_j^c \right\|^2 + \left\| \sqrt{\sigma_i^c} - \sqrt{\sigma_j^c} \right\|_2^2 \right)^{1/2}, \quad (1)$$

where μ is the mean of the average logits per class, σ is an approximate diagonal matrix of variances, and c is the class label. Further, we sort all Wasserstein distances to obtain the average logits of the top M clients, generally setting $M = 0.25 * K$. Thus, the synthetic logits ϕ_{syn} is computed as:

$$\phi_{syn} = \frac{1}{M} \frac{1}{\sqrt{\sum_{m=1}^{M} \phi_m^2}} \sum_{m=1}^{M} \phi_m. \quad (2)$$

Next, we perform logit calibration based on the average logits ϕ_k of client k and the synthetic logits ϕ_{syn}. The objective function for local logit calibration is expressed as:

$$\begin{aligned} &\mathcal{L}_{Local} = L_{KL}\left(softmax\left(\phi_k, t\right), softmax\left(\phi_{syn}, t\right)\right) \\ &\text{where} \quad softmax\left(z_i, t\right) = \frac{e^{z_i/t}}{\sum_{c=1}^{\mathcal{C}} e^{z_c/t}}, \end{aligned} \quad (3)$$

where L_{KL} is the Kullback-Leibler (KL) divergence between ϕ_k and ϕ_{syn}, z_i is the output value of the i-th dimension, and t represents the temperature value used to soften the logits.

Based on the $\mathcal{L}_{Local}$ objective function, the local model effectively alleviate local drift by transferring generalization knowledge. However, the long-tail effect of local data causes model predictions to be biased towards the head classes, and the local model cannot be effectively trained. Therefore, we propose the class-importance weighting strategy for class importance evaluation by the ratio between ϕ_{syn} and ϕ_k. During model training, the differences in the amount of data among various classes are balanced by assigning different weights to class samples. In particular, for the tail class samples, the loss is increased by class importance weighting so that the local model pays more attention. The form of class-importance weighting $\mathcal{L}_{IW}$ is denoted as follows:

$$\mathcal{L}_{IW} = \sum_{(x,y)\in\mathcal{D}^k} L_{CE}\left(w_k\left(x\right), y\right) \frac{\phi_{syn}}{\phi_k}, \tag{4}$$

where L_{CE} is the Cross-Entropy (CE) loss, x and y is the sample and label on client k, respectively.

In the local training phase, the local model is jointly trained based on $\mathcal{L}_{Local}$ and $\mathcal{L}_{IW}$ loss functions, which effectively solves the challenges brought by non-IID and long-tailed data.

Global Calibration Distillation. Although local logit calibration effectively alleviates the performance harm caused by the local model. However, when the universal class distribution is non-IID and long-tailed, the generalization ability of the global model on the tail classes and heterogeneous data is also poor. To address drift and long-tailed effects from a global perspective, we propose global calibration distillation on the server side. We first calibrate the logits of the global model based on the class-query strategy and then fully use the ensemble logits' generalization knowledge to distill the calibrated global logits further. Since the local models are trained on non-IID and long-tailed data, their prediction results are highly diverse, which is one of the important factors that make the ensemble model work better than a single model. Therefore, the ensemble logits contain more diverse and generalizable knowledge. First, the server receives the average logits from all participating clients, which is privacy-safe since the average logits cannot obtain prototype data information. Then, we weight all logits to get the ensemble logits ϕ_{ens}:

$$\phi_{ens} = \sum_{\mathrm{k}=1}^{K} \frac{|\mathcal{D}^k|}{|\mathcal{D}|} e_k \phi_k, \tag{5}$$

where e_k represents the ratio of the k-th client logits in the ensemble logits. We consider that e_k belongs to between 0 and 1, and the difference among different local models should be minimized, so e_k is defined as follows:

$$e_k = sigmoid\left(\phi_k \cdot r_k\right) + \lambda \left\|w_k\right\|^2, \tag{6}$$

where $r_k \in \mathbb{R}^{\mathcal{C}}$ is a learnable parameter, λ is used to control the degree of regularization, and $\cdot$ denotes the common matrix multiplication. We use the sigmoid function to normalize it between 0 and 1, while adding a regularization term to eliminate bias introduced by local model aggregation. Finally we normalize e_k by softmax to make its sum equal to 1.

Similar to FedAvg, we obtain the global model w_s by weighted averaging the parameters of all participating client models. To obtain the global logits, we utilize a small auxiliary dataset $\mathcal{D}_{sam}$ on the server, which is sampled from the training dataset. Since $\mathcal{D}_{sam}$ is obtained before training and stored on the server side, it does not lead to privacy leakage. Based on the global model w_s and $\mathcal{D}_{sam}$, we get the global logits ϕ_{glo}. However, if the local model cannot completely eliminate the negative influence of the long-tailed effect, the weighted global model is still biased towards the head classes. Therefore, we propose the class-query strategy, which uses learnable parameters a and b to calibrate the global logits according to different classes, further enhancing the confidence of the global model when querying the tail classes. The global logits further enhanced by the class-query strategy is denoted as:

$$\phi_{glo} = \sum_{c=1}^{\mathcal{C}} a_c \phi_{glo}^c + b_c, \tag{7}$$

where c represents the class label. The class-query performs class weight calibration on global logits by balancing the weights of head and tail classes.

We regard the ensemble logits with strong generalization ability as the teacher model, the global logits after class-query calibration as the student model, and then we perform calibration distillation. Specifically, we construct the following loss function based on calibration distillation:

$$\mathcal{L}_{Global} = L_{KL}\left(\frac{\phi_{glo}}{\sqrt{\left\|\phi_{glo}\right\|^2}}, \frac{\phi_{ens}}{\sqrt{\left\|\phi_{ens}\right\|^2}}\right). \tag{8}$$

Finally, the global model after calibration distillation can also use the auxiliary dataset $\mathcal{D}_{sam}$ to learn generalization knowledge from the ensemble logits, further improving the performance. We effectively alleviate the model training challenge with non-IID and long-tailed data through local calibration and calibration distillation in the local training stage and global aggregation stage.

IV. Experiments

In this section, we compare the performance of our proposed approach FedLC with other key related work. All results reported are the average of three repeating runs with a standard deviation of different random seeds, and are tested under the uniform settings.

A. Experimental Setup

FedLC is evaluated on EMNIST [31], CIFAR-10 and CIFAR-100 [30], and mini-ImageNet [32]. To simulate a scenario for FL with non-IID and long-tailed data, the training data in each dataset are split into multiple clients and the performance is evaluated on the test dataset.

EMNIST. The dataset consists of 47 different handwritten character and digit classes (image size 28). We first sample the auxiliary dataset $\mathcal{D}_{sam}$ from the entire training data, then reduce the number of training examples for each class according to [27], and keep the test set unchanged. Therefore,

TABLE I
RESULTS (TOP-1 TEST ACCURACY [%]) FOR EMNIST, CIFAR-10, AND CIFAR-100 IMAGE CLASSIFICATION OF DIFFERENT FL METHODS.

Methods	EMNIST			CIFAR-10			CIFAR-100		
	IF = 100	IF = 50	IF = 10	IF = 100	IF = 50	IF = 10	IF = 100	IF = 50	IF = 10
FedAvg	73.80	77.16	83.34	56.70	65.23	76.29	30.61	33.87	44.97
FedProx	73.16	77.59	83.22	56.85	64.96	75.82	30.75	34.15	45.08
FedBN	73.79	77.66	83.42	55.99	65.60	76.08	29.95	34.07	44.94
FedCurv	73.03	77.18	83.42	56.53	65.88	76.39	30.24	33.81	44.43
Scaffold	73.37	76.97	82.94	56.96	66.46	76.47	30.40	33.89	45.06
FedNova	72.97	76.36	83.61	56.72	66.14	76.30	30.35	34.01	44.61
FedBE	70.20	73.82	80.78	53.30	63.51	73.25	28.46	31.15	42.10
FedDF	70.64	74.57	81.30	52.53	62.90	73.04	27.93	31.64	42.35
FedGEN	70.31	74.63	81.55	53.82	62.97	73.27	28.04	31.17	42.63
FedLC (Ours)	**82.16**	**85.22**	**90.14**	**67.35**	**75.50**	**84.16**	**38.83**	**41.64**	**50.39**

the training data fits a long-tailed distribution with different imbalance factors (IF), which is calculated from the ratio between the number of samples in the largest class and that in the smallest class. Finally, on the basis of long-tailed data, we construct non-IID data with different parameter α based on Dirichlet distribution non-IID split method [33].

CIFAR-10/100. CIFAR-10 and CIFAR-100 datasets (image size 32) have a training set of 50,000 examples, and a test set of 10,000 examples. CIFAR-10 has 10 classes and CIFAR-100 has 100 classes. We remove the auxiliary dataset $\mathcal{D}_{sam}$, and then construct the heterogeneous long-tailed distribution with different IF and non-IID α.

mini-ImageNet. The mini-ImageNet dataset contains 100 categories, each category includes 500 training images, 50 validation images and 50 test images (image size 224). We first obtain the auxiliary dataset $\mathcal{D}_{sam}$, and then use the same strategy described above to create non-IID and long-tailed imbalance versions of mini-ImageNet.

Throughout the experiments, we report the performance of the FL global model. We use ResNet-8 for EMNIST, CIFAR-10 and CIFAR-100, and ResNet-18 [34] for mini-ImageNet as the backbone model. We run 200 global communication rounds, with 20 user models in total and active-user ratio at 40%. We adopt a local updating step $T = 10$, each step uses a mini batch with size $B = 64$, and SGD (learning rate 0.1) as the optimizer to update model. For the calibration process, in the local client, we set the number of selected clients $M = 5$, and the temperature parameter $t = 4.0$. In the global distillation phase, we set the hyperparameter λ at 0.001, the sampling rate of the auxiliary dataset at 5%, and the calibration distillation steps J at 100. Usually, we set the long-tailed ratio IF $= 100$ and the non-IID degree $\alpha = 0.4$.

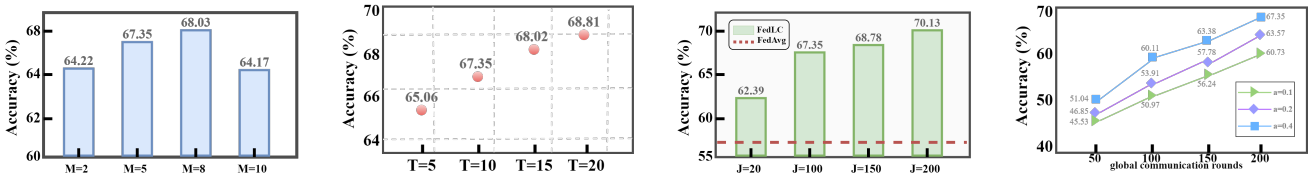
B. Performance Comparison

We compare our approach with recent state-of-the-art FL methods towards solving the non-IID and long-tailed problem. For local and global drifts, we compare the following FL methods: FedAvg [3], FedProx [9], FedBN [8], FedCurv [35], Scaffold [7] and FedNova [36], which constrain the heterogeneous dissimilarity by limiting local or global differences. Moreover, we also compare the distillation-based FL methods, including FedBE [37], FedDF [33] and FedGEN [38].

TABLE II
TOP-1 TEST ACCURACY (%) ON MINI-IMAGENET.

Methods	mini-ImageNet					
	IF = 100		IF = 50		IF = 10	
	$\alpha = 0.2$	$\alpha = 0.4$	$\alpha = 0.2$	$\alpha = 0.4$	$\alpha = 0.2$	$\alpha = 0.4$
FedAvg	30.81	32.50	36.66	38.27	44.07	46.25
FedProx	30.34	32.23	36.78	37.53	44.45	46.37
FedBN	30.31	32.83	36.89	38.36	44.38	46.77
FedCurv	30.39	32.16	36.35	38.22	43.86	45.87
Scaffold	30.66	31.95	36.29	38.07	44.31	46.21
FedNova	29.97	32.45	36.11	37.92	43.65	46.25
FedBE	25.19	27.06	32.07	34.58	40.91	42.15
FedDF	25.67	27.54	32.24	35.04	41.03	42.59
FedGEN	26.09	28.88	33.19	35.41	41.62	42.96
FedLC (Ours)	**37.61**	**40.69**	**42.47**	**44.39**	**50.22**	**51.27**

Results on EMNIST and CIFAR-10/100. The results (Top-1 accuracy in test dataset) on EMNIST, CIFAR-10 and CIFAR-100 are shown in Table I. Compared with the baseline FL methods, FedLC achieved a higher final accuracy (accuracy of the respective models after finishing all training rounds) across different datasets and imbalance levels. Especially in the setting of CIFAR-100 and IF $= 100$, the training data is complex and seriously unbalanced, but FedLC still achieved excellent performance, showing clearly the superiority of FedLC in dealing with non-IID and long-tailed data and its robust generalization. FedLC is robust against different imbalance levels of heterogeneous data while consistently performing well. No matter under which dataset, compared with the benchmark FedAvg, FedLC has the highest performance gain when IF $= 100$ (around 8.36% for EMNIST, around 10.65% for CIFAR-10, and 8.22% for CIFAR-100). FedProx, FedBN, FedCurv, Scaffold, and FedNova are similar to FedAvg because they only consider the data heterogeneity and ignore the adverse effects of long-tailed imbalance. However, FedLC addressed the non-IID issues from both local drift and global drift, and its performance is far superior to the above FL methods. In addition, we found that FedBE, FedDF, and FedGEN performed even worse than FedAvg. We consider that the strength of these distillation-based Fl methods is to distill local knowledge from the clients, thereby mitigating differences in potential distributions across different clients. Unfortunately, owing to the long-tailed effect, the local model



(a) Similar clients M. (b) Local update epochs T. (c) Calibration distillation steps J. (d) Non-IID degree α.

Fig. 4. Hyperparameter effects in FedLC.

does not perform well in predicting the long-tailed data, which further leads the distilled model to perform worse on the tail classes than the global model. Therefore, these methods perform worse than FedAvg because they do not consider the global imbalanced class distribution, which also confirms the necessity and effectiveness of the class-importance weighting strategy and the class-query strategy in FedLC. The advantages of FedLC in long-tailed applications are further demonstrated by local training weighting and learnable parameter optimization based on global calibration distillation.

Results on mini-ImageNet. The classification results for different IF and α on mini-ImageNet dataset are shown in Table II. From this result, we observe that FedLC significantly outperforms other methods with different heterogeneous long-tailed data on the mini-ImageNet dataset. It is obvious that other FL methods decrease seriously with more imbalance factor IF and fewer non-IID ratio α while our method declines slowly, which means that in the case of extremely heterogeneous and long-tailed data, our method FedLC may be more effective. This result also confirms the effectiveness of FedLC in decoupling heterogeneous data and long-tailed distribution through local logit calibration and global calibration distillation. In addition, for other FL methods, the impact of long-tailed distribution is more significant than non-IID, but FedLC effectively alleviates the damage of the long-tailed effect.

C. Analysis for FedLC

In this subsection, we further conduct hyperparameter and ablation studies based on the CIFAR-10 image classification to investigate key properties of our FedLC and the importance of each component. We uniformly set the non-IID degree $\alpha = 0.4$ and the imbalance factor IF $= 100$.

Hyperparameters setting. We mainly explore the following hyperparameters: for the local logit calibration stage, the number of similar clients M and different local update epochs T; for the global calibration distillation stage, the calibration distillation steps J. As shown in Fig. 4(a), as the number of similar clients M increases, the FedLC performance is also gradually improved. However, if M is too large, it may cause the synthetic logits to contain noise, which will degrade the performance of FedLC. Regarding the effect of different local update epochs T on FedLC, as shown in Fig. 4(b). We discovered that T affects the local logit calibration, and the larger T is, the better the average logits generated by the local model, further improving the FedLC performance. For the global calibration distillation stage, a larger number of distillation steps J helps the FL global model learn generalization knowledge from the ensemble logits, so the performance of FedLC increases steadily with the increase of J. This intuition is empirically confirmed by the experiments presented in Fig. 4(c). In addition, we also explored the impact of data heterogeneity α on FedLC, and the results are shown in Fig. 4(d). FedLC can still achieve better results when the data distributions are highly heterogeneous (with a small α), which also shows that FedLC is robust against different levels of data heterogeneity while consistently performing well.

TABLE III
ABLATION RESULTS ON CIFAR-10.

FedAvg	$\mathcal{L}_{Local}$	$\mathcal{L}_{Global}$	Δ_{CIW}	Δ_{CQ}	CIFAR-10
✓					56.70
✓	✓	✓			60.37
✓	✓	✓	✓		62.93
✓	✓	✓		✓	64.59
✓	✓	✓	✓	✓	**67.35**

Ablation study. To demonstrate the effect of different components of FedLC, we designed ablation experiments and split FedLC into four components: local logit calibration $\mathcal{L}_{Local}$, global calibration distillation $\mathcal{L}_{Global}$, class-importance weighting strategy Δ_{CIW}, and class-query strategy Δ_{CQ}, as shown in Table III. Without considering the long-tailed effect (no Δ_{CIW} and Δ_{CQ}), our method also performed better than FedAvg, which demonstrated the effectiveness of local and global calibration for drift issues. Further, FedLC achieved the best performance when using class-importance and class-query to alleviate local and global class imbalances, respectively. Compared with class-importance, class-query brings a greater performance gain, because the class-query strategy focuses more on distilling globally relevant class distribution knowledge from the ensemble logits.

V. CONCLUSION

In this paper, we proposed FedLC, a novel federated learning algorithm based on local and global logit calibration for solving the joint problem of heterogeneous and long-tailed data in FL. Our solution focused on decoupling data heterogeneity and long-tail effect, and alleviating the harm of heterogeneous long-tailed data through local logit calibration and global calibration distillation at both client and server perspectives. In particular, we proposed class-importance weighting and class-query strategies to address the imbalanced class distribution locally and globally. Meanwhile, logit calibration also effectively solved the drift issues in

the local and global model updates. We conducted extensive experiments and consistently demonstrated the effectiveness of FedLC, especially when the data is extremely heterogeneous and long-tailed. In the future, we will plan to apply FedLC to more personalized and flexible edge applications to solve the FL long-tailed problem. Overall, our work is helpful to promote the development of FL in real world applications by correcting biased logits on the client and server.

ACKNOWLEDGMENT

This work has been supported by National Natural Science Foundation of China under Grant No.61702398, No.61672415, and No.61872282, and the Fundamental Research Funds for the Central Universities under Grant No.JB211504.

REFERENCES

[1] P. P. Shinde and S. Shah, "A review of machine learning and deep learning applications," in *2018 International Conference on Computing Communication Control and Automation*, 2018, pp. 1–6.

[2] P. Voigt and A. Von dem Bussche, "The eu general data protection regulation (gdpr)," *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, vol. 10, no. 3152676, pp. 10–5555, 2017.

[3] H. B. McMahan, E. Moore, D. Ramage *et al.*, "Communication-efficient learning of deep networks from decentralized data," in *2017 International Conference on Artificial Intelligence and Statistics*, 2017, pp. 1273–1282.

[4] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Transactions on Intelligent Systems and Technology*, vol. 10, no. 2, pp. 1–19, Jan 2019.

[5] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Processing Magazine*, vol. 37, pp. 50–60, May 2020.

[6] Q. Li, Y. Diao, Q. Chen, and B. He, "Federated learning on non-iid data silos: An experimental study," in *2022 IEEE 38th International Conference on Data Engineering (ICDE)*, 2022, pp. 965–978.

[7] S. P. Karimireddy, S. Kale, M. Mohri *et al.*, "SCAFFOLD: Stochastic controlled averaging for federated learning," in *Proceedings of the 37th International Conference on Machine Learning*, vol. 119. PMLR, Jul 2020, pp. 5132–5143.

[8] X. Li, M. Jiang, X. Zhang, M. Kamp, and Q. Dou, "Fedbn: Federated learning on non-iid features via local batch normalization," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/pdf?id=6YEQUn0QICG

[9] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi *et al.*, "Federated optimization in heterogeneous networks," in *Proceedings of Machine Learning and Systems*, vol. 2, Apr 2020, pp. 429–450.

[10] C. T Dinh, N. Tran, and J. Nguyen, "Personalized federated learning with moreau envelopes," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 394–21 405, Jun 2020.

[11] A. Shamsian, A. Navon, E. Fetaya, and G. Chechik, "Personalized federated learning using hypernetworks," in *International Conference on Machine Learning*. PMLR, Mar 2021, pp. 9489–9502.

[12] L. Yang, H. Jiang, Q. Song, and J. Guo, "A survey on long-tailed visual recognition," *International Journal of Computer Vision*, pp. 1–36, 2022.

[13] Y. Zhang, B. Kang, B. Hooi, S. Yan, and J. Feng, "Deep long-tailed learning: A survey," *CoRR*, vol. abs/2110.04596, 2021. [Online]. Available: https://arxiv.org/abs/2110.04596

[14] H.-P. Chou, S.-C. Chang, J.-Y. Pan, W. Wei, and D.-C. Juan, "Remix: Rebalanced mixup," in *European Conference on Computer Vision*. Springer, 2020, pp. 95–110.

[15] S. Zhang, Z. Li, S. Yan, X. He, and J. Sun, "Distribution alignment: A unified framework for long-tail visual recognition," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 2361–2370.

[16] X. Ran, L. Ge, and L. Zhong, "Dynamic margin for federated learning with imbalanced data," in *2021 International Joint Conference on Neural Networks*, 2021, pp. 1–8.

[17] X. Shuai, Y. Shen, S. Jiang, Z. Zhao *et al.*, "Balancefl: Addressing class imbalance in long-tail federated learning," in *2022 21st ACM/IEEE International Conference on Information Processing in Sensor Networks*, 2022, pp. 271–284.

[18] Z. Chen, S. Liu, H. Wang, H. H. Yang *et al.*, "Towards federated long-tailed learning," *CoRR*, vol. abs/2206.14988, 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2206.14988

[19] M. J. Sheller, G. A. Reina, B. Edwards, J. Martin, and S. Bakas, "Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation," in *International MICCAI Brainlesion Workshop*. Springer, Oct 2018, pp. 92–104.

[20] A. Reisizadeh, F. Farnia, R. Pedarsani, and A. Jadbabaie, "Robust federated learning: The case of affine distribution shifts," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21 554–21 565, 2020.

[21] H. Wang, M. Yurochkin, Y. Sun, D. Papailiopoulos, and Y. Khazaeni, "Federated learning with matched averaging," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=BkluqlSFDS

[22] D. Jiang, C. Shan, and Z. Zhang, "Federated learning algorithm based on knowledge distillation," in *2020 International Conference on Artificial Intelligence and Computer Engineering*, 2020, pp. 163–167.

[23] J. Mills, J. Hu, and G. Min, "Multi-task federated learning for personalised deep neural networks in edge computing," *IEEE Transactions on Parallel and Distributed Systems*, vol. 33, no. 3, pp. 630–641, 2022.

[24] B. Zhou, Q. Cui, X.-S. Wei, and Z.-M. Chen, "BBN: Bilateral-branch network with cumulative learning for long-tailed visual recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9719–9728.

[25] B. Kang, S. Xie, M. Rohrbach, Z. Yan *et al.*, "Decoupling representation and classifier for long-tailed recognition," in *International Conference on Learning Representations*, 2020. [Online]. Available: https://openreview.net/forum?id=r1gRTCVFvB

[26] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9260–9269.

[27] K. Cao, C. Wei, A. Gaidon, N. Arechiga, and T. Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," *Advances in neural information processing systems*, vol. 32, 2019.

[28] M. Luo, F. Chen, D. Hu, Y. Zhang *et al.*, "No fear of heterogeneity: Classifier calibration for federated learning with non-iid data," *Advances in Neural Information Processing Systems*, vol. 34, pp. 5972–5984, 2021.

[29] X. Shang, Y. Lu, Y.-m. Cheung, and H. Wang, "Fedic: Federated learning on non-iid and long-tailed data via calibrated distillation," *arXiv preprint arXiv:2205.00172*, 2022.

[30] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images," 2009.

[31] G. Cohen, S. Afshar, J. Tapson, and A. Van Schaik, "Emnist: Extending mnist to handwritten letters," in *2017 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2017, pp. 2921–2926.

[32] O. Vinyals, C. Blundell, T. Lillicrap *et al.*, "Matching networks for one shot learning," in *Advances in Neural Information Processing Systems*, vol. 29. Curran Associates, Inc., 2016.

[33] M. Yang, X. Wang, H. Zhu, H. Wang, and H. Qian, "Federated learning with class imbalance reduction," in *2021 29th European Signal Processing Conference*, 2021, pp. 2174–2178.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[35] N. Shoham, T. Avidor, A. Keren, N. Israel *et al.*, "Overcoming forgetting in federated learning on non-iid data," *arXiv preprint arXiv:1910.07796*, 2019.

[36] J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, "Tackling the objective inconsistency problem in heterogeneous federated optimization," *Advances in neural information processing systems*, vol. 33, pp. 7611–7623, 2020.

[37] H. Huang, F. Shang, Y. Liu, and H. Liu, "Behavior mimics distribution: Combining individual and group behaviors for federated learning," *international joint conference on artificial intelligence*, 2021.

[38] Z. Zhu, J. Hong, and J. Zhou, "Data-free knowledge distillation for heterogeneous federated learning," in *International Conference on Machine Learning*. PMLR, 2021, pp. 12 878–12 889.