FEDIC: FEDERATED LEARNING ON NON-IID AND LONG-TAILED DATA VIA CALIBRATED DISTILLATION

Xinyi Shang[1], *Yang Lu*[1,†], *Yiu-ming Cheung*[2], *Hanzi Wang*[1]

[1]Xiamen University, [2]Hong Kong Baptist University
[1]shangxinyi@stu.xmu.edu.cn, [1]{luyang, hanzi.wang}@xmu.edu.cn, [2]ymc@comp.hkbu.edu.hk

2022 IEEE International Conference on Multimedia and Expo (ICME) | 978-1-6654-8563-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICME52920.2022.9860009

ABSTRACT

Federated learning provides a privacy guarantee for generating good deep learning models on distributed clients with different kinds of data. Nevertheless, dealing with non-IID data is one of the most challenging problems for federated learning. Researchers have proposed a variety of methods to eliminate the negative influence of non-IIDness. However, they only focus on the non-IID data provided that the universal class distribution is balanced. In many real-world applications, the universal class distribution is long-tailed, which causes the model seriously biased. Therefore, this paper studies the joint problem of non-IID and long-tailed data in federated learning and proposes a corresponding solution called Federated Ensemble Distillation with Imbalance Calibration (FEDIC). To deal with non-IID data, FEDIC uses model ensemble to take advantage of the diversity of models trained on non-IID data. Then, a new distillation method with logit adjustment and calibration gating network is proposed to solve the long-tail problem effectively. We evaluate FEDIC on CIFAR-10-LT, CIFAR-100-LT, and ImageNet-LT with a highly non-IID experimental setting, in comparison with the state-of-the-art methods of federated learning and long-tail learning. Our code is available at https://github.com/shangxinyi/FEDIC.

Index Terms— Federated learning, Non-IID, Long-tailed learning, Distillation

1. INTRODUCTION

In recent years, an increasing number of deep learning techniques have been deployed in mobile devices to handle data from different sources, e.g., cameras, microphones, GPS, and other sensors. These data play a key role in generating strong predictive models to provide better services to the users. However, transmitting user data to the server would bring high privacy risks for both the service providers and mobile users [1]. Recently, federated learning has received increasing attention due to its capacity for distributed machine learning with privacy protection [2]. Data privacy is guaranteed by storing data and training model locally on each client. The global model on the server is produced by aggregating local models transmitted from clients without the requirement of any data from them [3]. However, data heterogeneity is still a major challenge in federated learning. Since the data in each client may be drawn from different distributions without meeting the requirement of IID [4], training on this kind of data results in poor generalization ability of the global model.

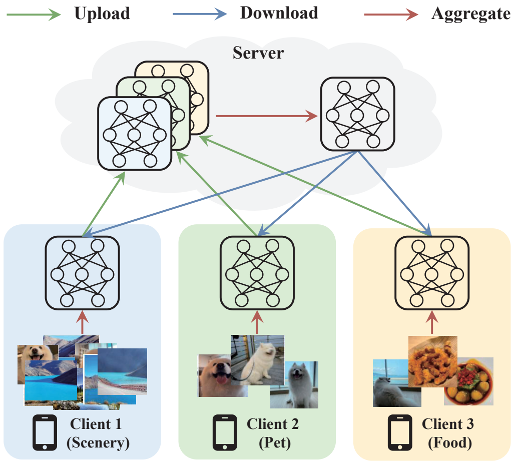

Fig. 1: An example of federated learning application for the task of gallery tagging on mobile phones. The majority class in each client is shown in parentheses.

In the literature, a number of methods have been proposed to deal with non-IID data in federated learning. They can be roughly categorized into client-side methods and server-side methods, respectively. The former aims to improve the local training process. Most of them regularize the local training process such that the diversity of client models can be limited [5]. The latter adopts specific model aggregation mechanisms to alleviate the negative influence of data heterogeneity [6, 7]. Some recent works have adopted knowledge distillation on the server [8, 9]. The knowledge is transfered from an ensemble model, which is built by local models, to the global model.

[†]Corresponding author: Yang Lu, luyang@xmu.edu.cn

This work was supported in part by the National Natural Science Foundation of China under Grants 62002302, 61872307, 61672444 and U21A20514; in part by the Open Research Projects of Zhejiang Lab (NO. 2021KB0AB03); in part by the National Natural Science Foundation of China/RGC Joint Research Scheme under Grant N_HKBU214/21; in part by the Natural Science Foundation of Fujian Province under Grant 2020J01005.

There are also some other methods that focus on optimization strategy [10, 11] on the server.

Although the abovementioned methods solve the data non-IIDness problem to some extent, they generally assume that the universal class distribution is balanced, which may not be true from a practical perspective. As shown in Fig. 1, if we consider the overall clients, a few classes like scenery have a large number of samples, while many classes like pet or food only take a small portion. Building a classification model on this kind of distribution is termed long-tail learning [12], which has been extensively studied in recent years. Some methods origin from the traditional imbalance learning [13], where the re-sampling [14] or re-weighing [15] techniques are adopted to alleviate the imbalance influence. The other methods [16] take advantage of the uniqueness of the deep learning model and focus on the representation learning.

The existing solutions for non-IID data in federated learning generally perform poorly on the tail classes due to the lack of consideration of the universal long-tail distribution. The global class distribution is long-tailed such that each client only holds a few tail classes, which makes local models perform poorly on the tail classes. Therefore, the global model aggregated by biased local models is also biased. There are also some methods specifically designed for federated learning on imbalanced data. One strategy is to adopt client selection to match complementary clients [17]. However, some clients may lose the chance to participate in model aggregation on the server if they cannot be matched with other clients. Recently, ratio loss [18] has been proposed to estimate the global imbalance status to help improve local optimization. However, the performance of ratio loss is dropped as the degree of data non-IIDness increases.

In this paper, we study the problem of federated learning on non-IID and long-tailed data, and correspondingly propose an effective server-side method called Federated Ensemble Distillation with Imbalance Calibration (FEDIC) without prior knowledge of global class distribution. In FEDIC, the knowledge distillation technique is adopted on the server to transfer the knowledge from the ensemble model to the global model. However, in the long-tailed setting, the ensemble model may still be biased towards the head classes. Subsequently, the transferred knowledge may not be helpful. We therefore propose a novel ensemble calibration method to eliminate the bias of the ensemble model before conducting knowledge distillation. Specifically, we first propose a new logit adjustment to reconstruct the ensemble model from the perspectives of clients and classes, respectively. Then, a calibration gating network is proposed to fuse the adjusted logits based on ensemble representations effectively. The final ensemble model generalizes well on both head and tail classes after calibration. The contributions of this paper can be summarized as follows:

- This paper is a pioneering work in federated learning to study the joint problems of the data non-IIDness and long-tail learning.
- We propose a new ensemble calibration method by logit adjustment and calibration gating network techniques to effectively make the output of the ensemble model unbiased.
- We propose an effective server-side federated learning method FEDIC, which utilizes the knowledge distillation technique to enhance the robustness of the global model on both non-IID and long-tailed data.

2. PROPOSED METHOD

In this section, we first describe the problem setting with some basic notations and then introduce FEDIC for federated learning on non-IID and long-tailed data.

2.1. Problem Setting

In this paper, the learning scenario is based on a typical federated learning system with K clients holding potentially non-IID local datasets $\mathcal{D}^1, \mathcal{D}^2, ..., \mathcal{D}^K$, respectively. The goal is to obtain a global model on the server over the union of all these datasets $\mathcal{D} \triangleq \bigcup_{k \leq K} \mathcal{D}^k$ without access to any data $\mathcal{D}^k$ on the k-th client. The setting difference in this paper is that $\mathcal{D}$ is drawn from a long-tailed distribution $(\mathcal{X}, \mathcal{Y})$, $\mathcal{Y} \in \{1, ..., C\}$, which is unknown in advance. The model in federated learning is typically a neural network $\phi_{\mathbf{w}}$ with parameters $\mathbf{w}$. $\phi_{\mathbf{w}}$ has two components: 1) a feature extractor $f_{\mathbf{w}}$, mapping each sample $\mathbf{x}$ to a d-dim representation vector; 2) a classifier $h_{\mathbf{w}}$, typically being a fully-connected layer which outputs logits to denote class confidence scores. The parameters of client k's local model are denoted as $\mathbf{w}_k$.

2.2. Proposed Method

FEDIC is a server-side method based on FedAvg [3] without intervening in the local training process on each client. It is based on an intuitive idea: The ensemble of the local client models has better generalization ability than the global model produced by parameter averaging [8]. Since the local models are trained on non-IID data, their prediction results are highly diverse, which is one of the most important factors that make the ensemble model work better than a single model [19]. However, due to model heterogeneity, the ensemble model cannot be transmitted to the clients for further updating. Therefore, it is straightforward to leverage knowledge distillation [20] to transfer the generalization ability from the ensemble model to the global model. Then, the distilled global model is transmitted to each client for further updating.

Specifically, on the server, we can construct the ensemble model as the teacher model:

$$\phi^t(\mathbf{x}) = \sum_{k=1}^{K} e_k \phi_{\mathbf{w}_k}(\mathbf{x}), \tag{1}$$

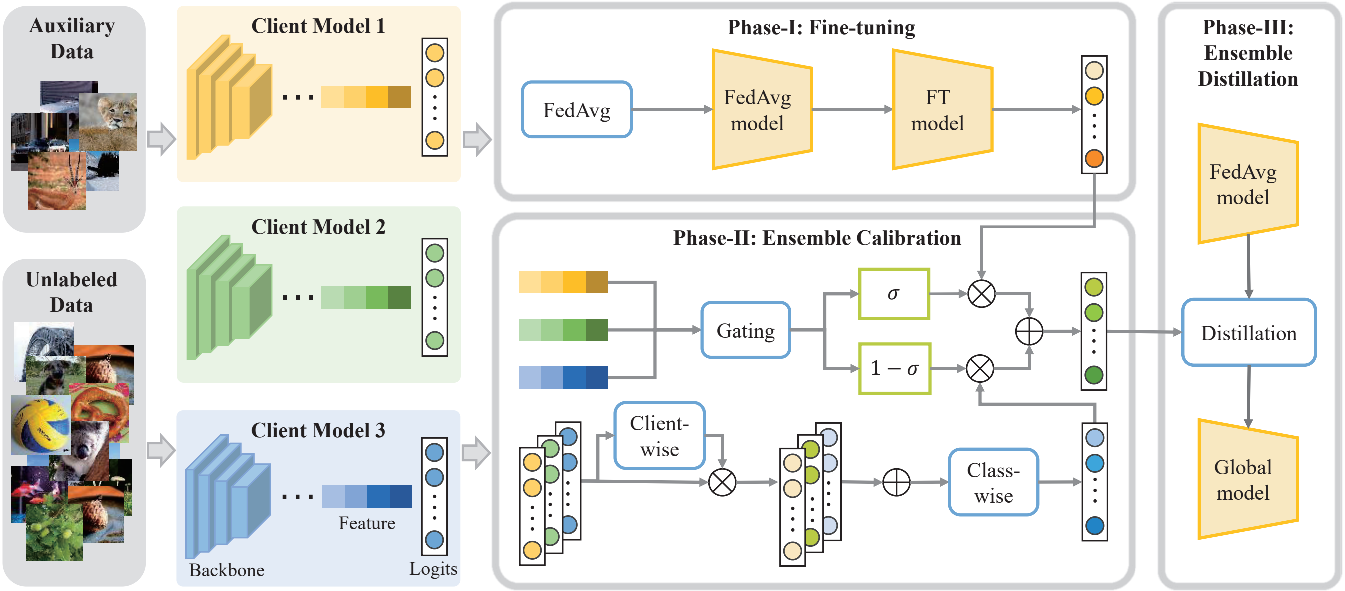

Fig. 2: The framework of FEDIC on the server.

where e_k is the ensemble weight for client k's local model. Then, we can obtain the global model $\mathbf{w}$ as the corresponding student model [1] [3]:

$$\mathbf{w} = \sum_{k=1}^{K} \frac{|\mathcal{D}^k|}{|\mathcal{D}|} \mathbf{w}_k, \tag{2}$$

$$\phi^s(\mathbf{x}) = \phi_{\mathbf{w}}(\mathbf{x}). \tag{3}$$

However, when the universal class distribution is long-tailed, the generalization ability of the ensemble model on the tail classes is also poor. As a result,it cannot provide a helpful guide to the student model on the tail classes. Therefore, we propose to utilize a small auxiliary dataset $\mathcal{D}_{aux}$ on the server, which is labeled and balanced in order to calibrate the ensemble model against the long-tail distribution. The main reason of utilizing the auxiliary data is that the global imbalance degree is unknown for both the server and clients, which makes most of the methods for long-tailed learning are infeasible. It is worth noting that all auxiliary data is collected independently on the server and there is no data transmission or data sharing in our problem setting. The model architecture of FEDIC on the server is shown in Fig. 2. In the following, we will describe two core components of FEDIC in detail.

Ensemble Calibration. Because of training on local data with different distributions, each local model may perform differently on the tail classes. It is reasonable to assign a higher ensemble weight to the local model that performs well on the tail classes to improve the generalization ability of the ensemble model. However, the server has no prior knowledge of which classes are the tail classes and which local model performs well on them. Therefore, instead of giving each client a static weight (e.g., $1/K$ for the common average ensemble) in the ensemble, we propose the client-wise logit adjustment that searches proper ensemble weights $e_k, k = 1, ..., K$ by learnable parameters. Given a sample $\mathbf{x} \in \mathcal{D}_{aux}$ on the server, we first calculate the logits of local models $\phi_{\mathbf{w}_k}(\mathbf{x})$. The ensemble weights e_k are calculated by a non-linear transform:

$$e_k = \text{sigmoid}\big(\mathbf{a}_e^T \phi_{\mathbf{w}_k}(\mathbf{x}) + b_e\big), \tag{4}$$

where $\mathbf{a}_e \in \mathbb{R}^C$ and b_e is a learnable parameter. e_k is then normalized to make its sum equal to 1. Subsequently, the weighted logits of the local model can be computed, as shown in Eq. (1). However, if none of the clients handles the tail classes well, the weighted ensemble is still biased towards the head classes. Subsequently, we propose class-wise logit adjustment to further enhance the logit of the tail classes by learnable parameters $\mathbf{a}_z, \mathbf{b}_z \in \mathbb{R}^C$. They linearly transform the original weighted ensemble logits $\phi^t(\mathbf{x})$ to calibrated logits $\mathbf{z}^{cl}$ on each class:

$$\mathbf{z}^{cl} = \mathbf{a}_z \odot \phi^t(\mathbf{x}) + \mathbf{b}_z, \tag{5}$$

where $\odot$ denotes the Hadamard product. Thus, $\mathbf{z}^{cl}$ is the calibrated logits after client-wise and class-wise logit adjustment.

However, the effectiveness of the premise of logit adjustment is that the features are well extracted. Simply manipulating the logits may not be sufficient if the feature extractors of local models are severely affected. Therefore, we propose to update the feature extractor as well to complement logit adjustment. Specifically, we can obtain a model $\widehat{\mathbf{w}}$ by fine-tuning the global model on $\mathcal{D}_{aux}$. Since $\mathcal{D}_{aux}$ is balanced, $\widehat{\mathbf{w}}$ is adjusted to obtain an unbiased feature extractor. Then, we can obtain the fine-tuned logits $\mathbf{z}^{ft} = \phi_{\widehat{\mathbf{w}}}(\mathbf{x})$ for the input $\mathbf{x}$. The logits $\mathbf{z}^{cl}$ and $\mathbf{z}^{ft}$ are both adjusted to deal with the long-tail distribution but they are from different perspectives. That

[1] For better notation representation, we ignore the superscript (t) to denote the model in the t-th round, which is usually adopted in federated learning literature.

Table 1: Top-1 test accuracy (%) for FEDIC and compared FL methods on CIFAR-10/100-LT with different IFs.

Family	Method	CIFAR-10-LT			CIFAR-100-LT		
		IF=100	IF=50	IF=10	IF=100	IF=50	IF=10
FL methods	FedAvg	52.12	52.43	59.97	25.81	28.19	38.22
	FedAvgM	53.64	54.42	59.52	25.11	28.82	38.77
	FedProx	52.75	55.07	60.44	25.43	27.77	38.45
	FedNova	52.93	56.53	61.58	26.81	28.91	39.62
Distillation-based FL methods	FedDF	50.33	52.58	58.84	25.60	28.79	38.60
	FedBE	44.05	50.66	53.53	22.46	23.77	33.53
Imbalance-oriented FL methods	Fed-Focal Loss	49.66	52.02	59.68	24.66	26.04	35.54
	Ratio Loss	54.15	57.77	60.58	26.72	28.83	38.79
	FedAvg+cRT	51.74	55.87	61.11	30.73	31.47	39.75
	FedAvg+τ-norm	44.38	45.59	48.29	19.59	22.07	30.48
	FedAvg+LWS	44.48	46.20	55.17	20.70	23.24	32.31
Proposed method	FEDIC	**63.11**	**63.82**	**65.50**	**33.67**	**34.74**	**41.93**

is, $\mathbf{z}^{cl}$ is produced on the merit of the model ensemble but with fixed feature extractors, while $\mathbf{z}^{ft}$ is based on the single global model, but its feature extractor is fine-tuned on $\mathcal{D}_{aux}$. Inspired by [21], we propose a calibration gating network to control the trade-off between $\mathbf{z}^{ft}$ and $\mathbf{z}^{cl}$, in order to effectively integrate the calibrated and fine-tuned logits and make them complement each other. The network takes the feature ensemble as the input through a non-linear layer to output the weight between $\mathbf{z}^{ft}$ and $\mathbf{z}^{cl}$, such that each sample obtains a different weight according to its own feature. The calibration gating network is formulated as:

$$\sigma = \text{sigmoid}(\mathbf{u}^T\mathbf{v}), \tag{6}$$

where $\mathbf{v} = \frac{1}{|S|}\sum_{k\in S} f_{\mathbf{w}_k}(\mathbf{x})$ is the feature ensemble, and $|S|$ is the number of selected clients in each round. $\mathbf{u} \in \mathbb{R}^d$ is a learnable parameter. Thus, the final calibrated logits $\mathbf{z}'$ through the calibration gating network is formulated as:

$$\mathbf{z}' = \sigma\mathbf{z}^{cl} + \big(1-\sigma\big)\mathbf{z}^{ft}. \tag{7}$$

The weight $\sigma \in (0, 1)$ acts as a feature-dependent gate to control the trade-off between $\mathbf{z}^{ft}$ and $\mathbf{z}^{cl}$. All learnable parameters in the whole process of ensemble calibration are updated by cross-entropy loss on $(\mathbf{x}, \mathbf{y}) \sim \mathcal{D}_{aux}$:

$$L = -\sum_{j=1}^{C} y_j \log \frac{\exp(z'_j)}{\sum_{i=1}^{C}\exp(z'_k)}. \tag{8}$$

Ensemble Distillation. To better distill unbiased knowledge from the teacher model (i.e., the calibrated ensemble model) to the student model (i.e., the global model), we follow the work of knowledge distillation with two loss components [20]: (1) L_{CE} is the cross-entropy loss between logits of the student model and the ground truth; (2) L_{KL} is the Kullback-Leibler (KL) divergence of the logits between the teacher model and the student model. We use $\mathcal{D}_{aux}$ to calculate L_{CE}, and use another unlabeled dataset $\mathcal{D}_{ulb}$ for L_{KL} to boost the performance of distillation knowledge further. Thus, the loss is constituted by a trade-off hyperparameter $\lambda \in [0, 1]$:

$$L_{FEDIC} = (1-\lambda)L_{CE} + \lambda L_{KL}. \tag{9}$$

We set $\lambda = 0.5$ in all experiments.

3. EXPERIMENTS

3.1. Experiment Setup

We conduct the experiments on the following datasets:

CIFAR-10/100-LT [22]. We first exclude the auxiliary data $\mathcal{D}_{aux}$ from the training data and then follow [23] to shape the rest of the data into a long-tail distribution with different imbalance factors (IF), which is calculated by the ratio between the number of samples in the largest class and that in the smallest class. For the unlabeled dataset $\mathcal{D}_{ulb}$, we use CIFAR-100 for CIFAR-10-LT, and use the downsampled ImageNet (image size 32) for CIFAR-100-LT.

ImageNet-LT is a long-tailed version of ImageNet [24]. It contains 115.8K images from 1,000 categories, with the largest and smallest categories containing 1,280 and 5 images, respectively. We obtain the auxiliary data $\mathcal{D}_{aux}$ from the balanced evaluation data and we use the oversampled CIFAR100 (image size 224) as $\mathcal{D}_{ulb}$.

We use ResNet-8 for CIFAR-10-LT and CIFAR-100-LT, and ResNet-50 for ImageNet-LT as the backbone network. By default, we run 200 global communication rounds, with 20 clients in total and an active user ratio $C = 40\%$ in each round. For local training, the batch size is set at 128 with learning rate 0.1 and SGD as the optimizer. For server training, we set the calibration steps I at 100, the distillation steps J at 100, and Adam with a learning rate 0.001 is used for knowledge distillation. Following [8], we use Dirichlet distribution to generate the non-IID data partition among clients with the concentration parameter $\alpha = 0.1$.

3.2. Comparison with the State-of-the-art Methods

To verify the effectiveness of FEDIC, we compare the proposed method with the following federated learning (FL) methods: FedAvg [3], FedAvgM [11], FedProx [5] and FedNova [10], and distillation-based FL methods, including FedDF [8] and FedBE [9]. All of them aim at producing a good global model on non-IID data. Moreover, we also compare the imbalance-oriented FL methods: Fed-Focal Loss [25], Ratio Loss [18], and FedAvg with post-hoc methods like cRT , τ-norm and LWS [16].

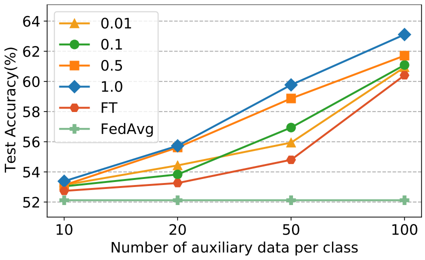

Fig. 3: The performance of FEDIC with different sizes of auxiliary data on CIFAR-10-LT with IF=100.

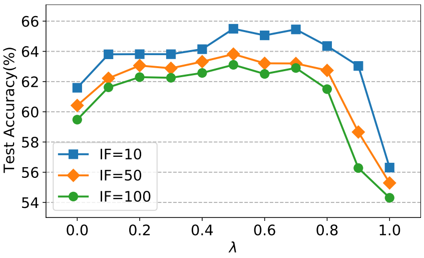

Fig. 4: The performance of FEDIC with different values of λ on CIFAR-10-LT.

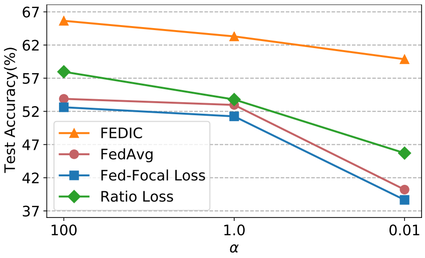

Fig. 5: The performance of FEDIC with different degrees of non-IIDness on CIFAR-10-LT with IF=100.

Table 2: Top-1 test accuracy (%) for FEDIC and compared FL methods on ImageNet-LT.

Method	ImageNet-LT			
	All	Many	Medium	Few
FedAvg	23.85	34.92	19.18	7.41
FedAvgM	22.57	33.93	18.55	6.73
FedProx	22.99	34.25	17.06	6.37
FedDF	21.63	31.78	15.52	4.48
Ratio Loss	24.32	36.33	18.14	7.10
FedAvg+LWS	21.58	31.66	15.76	5.33
FEDIC	**28.93**	**38.24**	**25.28**	**15.91**

Table 3: Ablation study on the components of ensemble calibration in FEDIC on CIFAR-10-LT with IF=100.

	Module			CIFAR-10-LT			
	FT	LA	KD	All	Many	Medium	Few
(a)	✗	✗	✗	53.59	**76.90**	50.12	34.90
(b)	✓	✗	✗	61.11	64.30	62.47	56.10
(c)	✗	✓	✗	60.66	66.00	60.93	53.26
(d)	✓	✓	✗	**64.77**	65.26	**64.88**	**62.55**
(e)	✗	✗	✓	53.92	**77.20**	51.32	34.05
(f)	✓	✗	✓	60.18	64.68	58.04	56.32
(g)	✗	✓	✓	61.71	63.75	63.45	57.94
(h)	✓	✓	✓	**63.11**	64.90	**63.60**	**60.83**

Results on CIFAR-10/100-LT. The results are summarized in Table 1. FEDIC achieves the highest test accuracy on both datasets with different IFs. Compared with the baseline FedAvg, the performance gain of FEDIC is the highest when IF=100 (around 11% for CIFAR-10-LT and 7.8% for CIFAR-100-LT). It shows the generalization ability of FEDIC when the universal class distribution is highly long-tailed. FedAvgM, FedProx and FedNova perform similarly to FedAvg because they only deal with data non-IIDness without taking global imbalanced class distribution into account. For the distillation-based methods, FedDF and FedBE perform even worse than FedAvg. A plausible reason is that their effectiveness is based on the power of the ensemble model as the teacher model to transfer knowledge. However, the ensemble model may perform even worse than the global model on the tail classes leading to a worse distilled model due to the global imbalanced distribution. This observation also validates the necessity of ensemble calibration in FEDIC. For the imbalance-oriented FL methods, some of them (e.g., Ratio Loss) perform well in some cases compared with FedAvg. However, there is still a performance gap compared with FEDIC because they only alleviate the imbalance problem on the server but ignore the data non-IID problem.

Results on ImageNet-LT. We evaluate FEDIC on ImageNet-LT whose results are reported in Table 2. Compared with the other methods, FEDIC achieves the best results on all cases we have tried thus far. At the same time, the accuracy on the few-shot classes achieves 15.91%, which is a significant improvement of 8.5% in comparison with the baseline.

3.3. Model Validation

Ablation study on ensemble calibration. We conduct an ablation study to evaluate the necessity of each component of ensemble calibration in FEDIC, as shown in Table 3. We evaluate three modules: Fine-tune (FT), client-wise and class-wise logit adjustment (LA) and knowledge distillation (KD). Note that we do not specifically evaluate the proposed calibration gating network because it is used only if both FT and LA are activated. The experiment is done by running FEDIC to round 200 and evaluating all combinations on that round. In the upper part (a)-(d) in Table 3, we only evaluate the performance of the ensemble model without distillation. Compared with the baseline (a), the overall accuracy of the calibrated ensemble model (d) is improved by 11.2%. In the lower part (e)-(h), knowledge distillation is conducted. It can be observed that the gap of the overall accuracy between the teacher model (d) and the student model (h) is only 1.7%, which indicates that the generalization ability to deal with the long-tail distribution is successfully transferred.

Influence of sizes of auxiliary and unlabeled datasets. The sizes of $\mathcal{D}_{aux}$ and $\mathcal{D}_{ulb}$ play a key role in FEDIC. Therefore, we evaluate its influence on the performance of FEDIC, compared with the baseline FedAvg and a global model fine-tuned with $\mathcal{D}_{aux}$ (marked as FT), as shown in Fig. 3. Different curves in the figure indicate the data fractions of the unlabeled data used for distillation. We observe that FEDIC consistently outperforms FedAvg and FT for all sizes of $\mathcal{D}_{aux}$.

Sensitivity analysis of hyperparameters. We investigate the impact of distillation trade-off coefficient λ. This hyperparameter controls the strength of distillation in the loss function in Eq. (9). It can be observed from Fig. 4 that FEDIC is robust to most λ values. However, the performance severely drops when λ reaches 1, which shows that solely distillation with unlabeled data is not enough for a good global model.

Influence of the degree of non-IIDness. Fig. 5 further shows the test accuracy of four methods under the different degrees of non-IIDness. It can be observed that the performance of all methods drops as the degree of non-IIDness increases. However, the performance of the compared methods drops more severely than FEDIC when α decreases from 1.0 to 0.01.

4. CONCLUSION

In this paper, we have proposed FEDIC to deal with the problem of learning a global model on non-IID and long-tailed data in the federated learning framework. FEDIC is a server-side method that first calibrates the biased ensemble model against the long-tail distribution by client-wise and class-wise logit adjustment with a calibration gating network. Then, the calibrated ensemble is used as the teacher model to transfer knowledge to the global model for further optimization on the clients. Also, the effectiveness of each component in FEDIC has been validated empirically. Experiments have shown that FEDIC outperforms the state-of-the-art FL methods on datasets with the non-IID and long-tailed setting.

5. REFERENCES

[1] Chang Xia, Jingyu Hua, Wei Tong, Yayuan Xiong, and Sheng Zhong, "A privacy-preserving scheme for convolutional neural network-based applications in mobile cloud," in *ICME*, 2020, pp. 1–6.

[2] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 50–60, 2020.

[3] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *AISTATS*, 2017, pp. 1273–1282.

[4] 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.

[5] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith, "Federated optimization in heterogeneous networks," in *Mach. Learn.*, 2020, pp. 429–450.

[6] Yaoxin Zhuo and Baoxin Li, "Fedns: Improving federated learning for collaborative image classification on mobile clients," in *ICME*, 2021, pp. 1–6.

[7] Ching-Hao Wang, Kang-Yang Huang, Jun-Cheng Chen, Hong-Han Shuai, and Wen-Huang Cheng, "Heterogeneous federated learning through multi-branch network," in *ICME*, 2021, pp. 1–6.

[8] Tao Lin, Lingjing Kong, Sebastian U Stich, and Martin Jaggi, "Ensemble distillation for robust model fusion in federated learning," in *NeurIPS*, 2020, pp. 2351–2363.

[9] Hong-You Chen and Wei-Lun Chao, "Fedbe: Making bayesian model ensemble applicable to federated learning," in *ICLR*, 2021, pp. 1–19.

[10] Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H. Vincent Poor, "Tackling the objective inconsistency problem in heterogeneous federated optimization," in *NeurIPS*, 2020, pp. 7611–7623.

[11] Tzu-Ming Harry Hsu, Hang Qi, and Matthew Brown, "Measuring the effects of non-identical data distribution for federated visual classification," *arXiv preprint arXiv:1909.06335*, 2019.

[12] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng, "Deep long-tailed learning: A survey," *arXiv preprint arXiv:2110.04596*, 2021.

[13] Haibo He and Edwardo A Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, 2009.

[14] Hsin-Ping Chou, Shih-Chieh Chang, Jia-Yu Pan, Wei Wei, and Da-Cheng Juan, "Remix: Rebalanced mixup," in *ECCV*, 2020, pp. 95–110.

[15] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng, "Meta-weight-net: Learning an explicit mapping for sample weighting," in *NeurIPS*, 2019, p. 1567–1578.

[16] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis, "Decoupling representation and classifier for long-tailed recognition," in *ICLR*, 2020, pp. 1–16.

[17] Moming Duan, Duo Liu, Xianzhang Chen, Renping Liu, Yujuan Tan, and Liang Liang, "Self-balancing federated learning with global imbalanced data in mobile systems," *IEEE Trans. Parallel Distrib. Syst.*, vol. 32, no. 1, pp. 59–71, 2020.

[18] Lixu Wang, Shichao Xu, Xiao Wang, and Qi Zhu, "Addressing class imbalance in federated learning," in *AAAI*, 2021, pp. 10165–10173.

[19] Thomas G Dietterich, "Ensemble methods in machine learning," in *Int. Workshop Multiple Classifier Syst.*, 2000, pp. 1–15.

[20] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[21] Songyang Zhang, Zeming Li, Shipeng Yan, Xuming He, and Jian Sun, "Distribution alignment: A unified framework for long-tail visual recognition," in *CVPR*, 2021, pp. 2361–2370.

[22] Krizhevsky Alex and Hinton Geoffrey, "Learning multiple layers of features from tiny images," in *Tech. Rep.* 2009, pp. 32–33, University of Toronto.

[23] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma, "Learning imbalanced datasets with label-distribution-aware margin loss," in *NeurIPS*, 2019, pp. 1567–1578.

[24] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, and Michael Bernstein, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[25] Dipankar Sarkar, Ankur Narang, and Sumit Rai, "Fed-focal loss for imbalanced data classification in federated learning," *arXiv preprint arXiv:2011.06283*, 2020.