Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Federated learning on non-IID and long-tailed data via dual-decoupling*

Zhaohui WANG, Hongjiao LI†‡, Jinguo LI†, Renhao HU, Baojin WANG

College of Computer Science and Technology, Shanghai University of Electric Power, Shanghai 201306, China

†E-mail: hjli@shiep.edu.cn; lijg@shiep.edu.cn

Received Apr. 23, 2023; Revision accepted Aug. 22, 2023; Crosschecked Mar. 1, 2024

Abstract: Federated learning (FL), a cutting-edge distributed machine learning training paradigm, aims to generate a global model by collaborating on the training of client models without revealing local private data. The co-occurrence of non-independent and identically distributed (non-IID) and long-tailed distribution in FL is one challenge that substantially degrades aggregate performance. In this paper, we present a corresponding solution called federated dual-decoupling via model and logit calibration (FedDDC) for non-IID and long-tailed distributions. The model is characterized by three aspects. First, we decouple the global model into the feature extractor and the classifier to fine-tune the components affected by the joint problem. For the biased feature extractor, we propose a client confidence re-weighting scheme to assist calibration, which assigns optimal weights to each client. For the biased classifier, we apply the classifier re-balancing method for fine-tuning. Then, we calibrate and integrate the client confidence re-weighted logits with the re-balanced logits to obtain the unbiased logits. Finally, we use decoupled knowledge distillation for the first time in the joint problem to enhance the accuracy of the global model by extracting the knowledge of the unbiased model. Numerous experiments demonstrate that on non-IID and long-tailed data in FL, our approach outperforms state-of-the-art methods.

Key words: Federated learning; Non-IID; Long-tailed data; Decoupling learning; Knowledge distillation
https://doi.org/10.1631/FITEE.2300284 **CLC number:** TP18

1 Introduction

An efficient distributed learning framework that addresses the issue of data silos without disclosing local private data is federated learning (FL). Classical FL, represented by the federated averaging (FedAvg) algorithm (McMahan et al., 2017), aggregates the local model uploaded by clients on the server side to produce a more comprehensive global model. Specifically, FL is divided into two steps—local training and global aggregation. During local training, the selected active clients update the global model downloaded from the server with their private data. The server then aggregates all uploaded local models using the FedAvg algorithm to regenerate the global model. This iterative process continues until the model converges. During the above process, only the parameters of the model are transmitted between the clients and the server. FL serves as an effective communication and privacy-preserving framework, demonstrating its ability to revolutionize real-world applications, including recommendation (Jalalirad et al., 2019; Tan et al., 2020), natural language processing (Jiang et al., 2021a), and medical treatment (Choudhury et al., 2019).

Along with its promising prospect, FL faces two data distribution scenarios that significantly affect its performance in the actual world. One scenario involves local data samples from different clients

‡ Corresponding author
* Project supported by the National Natural Science Foundation of China (No. 61702321)
ORCID: Zhaohui WANG, https://orcid.org/0009-0007-3676-209X; Hongjiao LI, https://orcid.org/0000-0003-0642-9046

that are non-independent and identically distributed (non-IID). The other important scenario is characterized by the global data distribution showing a long-tailed distribution, where both the number and the quality of the head classes severely outperform those of the tail classes. According to Wang LX et al. (2021), directly using the FedAvg algorithm on such non-IID and long-tailed data will seriously affect the accuracy of the aggregated model and skew the model heavily towards the head classes. As a result, dealing with non-IID and long-tailed data in FL is an extremely complicated and challenging task.

According to some recent studies (Li X et al., 2020; Zhao Y et al., 2022), using non-IID data for training models in FL can seriously impair the accuracy of the global model or potentially disrupt model convergence. Non-IID data not only create divergence in the distribution of the local and global data, but also lead to inconsistent local objective function and global optimization direction. To address the non-IID problem, a number of strategies have been put forth. These strategies are divided into two complementary perspectives. One perspective is to focus on local training by adding various regularization terms to prevent the local model from deviating too much from the global model (Karimireddy et al., 2020; Li T et al., 2020; Wang JY et al., 2020). The other perspective is to mitigate the impact of data heterogeneity by improving the model aggregation mechanism, such as using knowledge distillation (KD) (Li DL and Wang, 2019; Zhu et al., 2021), ensemble model (Lin T et al., 2020; Chen HY and Chao, 2021), and shared data (Fang and Ye, 2022). The most cutting-edge works (Li XC and Zhan, 2021; Luo et al., 2021) analyzed the reasons for performance degradation of FL on heterogeneous data in the structure of deep neural networks. Since the last classification layer is particularly vulnerable to non-IID distribution changes, the unintended consequences of biased classifiers on deep neural networks can be catastrophic. Long-tailed distribution revealed by global data is also problematic in our scenario.

In the actual world, data belonging to the head classes constitute a great proportion of the population, while data from the tail classes, generated by uncommon events, are scarce (He HB and Garcia, 2009). This particular class imbalance distribution has been defined as long-tailed distribution by researchers (Shen ZB et al., 2022). The effect of long-tailed distribution in FL setting is even more catastrophic, since a class that is a minority locally can actually be a majority class globally. Most of the early studies on long-tailed distribution followed the imbalanced learning method and focused on learning one-stage models, e.g., balanced sampling (Han et al., 2005; Pouyanfar et al., 2018) and re-weighting (Cui et al., 2019; Huang et al., 2020; Zhang et al., 2021; Alshammari et al., 2022). Nevertheless, it is difficult to use these methods directly in FL with long-tailed distribution because they require global class distribution. A few research works (Wang LX et al., 2021; Shen ZB et al., 2022) have recently started to concentrate on the class imbalance issue in FL. Although such approaches alleviate the joint problem to some extent, they often neglect the impact of inconsistency between local data and global data on the model. Solutions to both non-IID and long-tailed distribution are extremely concerned about the performance of the classifier. The reason is that the bottom layers of neural networks are used to extract features in cases where most objects have similar characteristics, whereas the topmost layer is used for the most specific tasks (Li XC and Zhan, 2021). However, the feature extractor, which is the main part of the model, has received little attention. Decoupling the training process into representation learning and classifier re-training is one of the most recent developments in long-tailed learning (Kang et al., 2020). That is, both the representation learning of the feature extractor and the decision boundary of the classifier deserve attention.

In this paper, we present federated dual-decoupling via model and logit calibration (FedDDC), a novel solution for addressing the joint issue of non-IID and long-tailed distribution in FL, inspired by the concept of decoupling learning. Our approach involves decoupling the global model into the feature extractor and classifier. Since aggregating a well-performing global model in our scenario is extremely difficult, we focus on learning a better representation by calibrating the feature extractors and shifting decision boundaries by adjusting the classifier. Specifically, we propose a new re-weighting method, namely, target-loss confidence re-weighting (TLCR), to calibrate the feature extractors, on one hand, and use classifier re-balancing (cRB) to adjust the biased classifier on the other hand. To integrate

and calibrate the models, we devise a useful strategy considering the diversity of models trained on non-IID data and the scarcity of tail classes in the long-tailed distribution. For this purpose, we introduce an adaptive calibration function that uses a nonlinear layer to balance the output logits from different models. Finally, we decouple not only the structure but also the output of the model. Since the ensemble logits contain both target class knowledge and non-target class knowledge, we use the latest decoupled KD in the joint problem for the first time to effectively distill knowledge from the ensemble model to the global model. The main contributions of this work are as follows:

1. We propose an FL framework, the FedDDC, which studies FL problems based on non-IID and long-tailed data from the perspectives of decoupling model and knowledge.

2. We propose a novel confidence-weighting method that efficiently reassigns the weights for each client through the loss of the client model and the logits of the target class.

3. We validate the proposed approach in a variety of contexts with different degrees of non-IIDness and imbalance factors. According to the results of the experiments, FedDDC routinely outperforms state-of-the-art FL methods in terms of performance.

2 Related works

2.1 FL with non-IID data

The well-known aggregation method in FL, FedAvg, often suffers from severe accuracy degradation when the data are of the non-IID type. To address the non-IID challenge in FL, a variety of solutions have been put forward. Some approaches rely on shared data (Fang and Ye, 2022), lifelong learning (Shoham et al., 2019), or the ensemble method (Lin T et al., 2020; Chen HY and Chao, 2021) to reduce the influence of non-IID-ness on the server. A popular method is adding regularization terms during local training (Karimireddy et al., 2020; Yu et al., 2020). FedProx (Li T et al., 2020) introduces a proximal term and limits too much divergence between the local and global model parameters. Taking advantage of KD (Li DL and Wang, 2019; Chen HY and Chao, 2021) is a better solution to adapt to the new task. Hinton et al. (2015) initially put out the concept of KD. KD aims to extract knowledge by distillation from a pre-trained network (i.e., teacher) to a network that requires training (i.e., student). Based on this foundation, KD is introduced to FL for knowledge transfer between clients and servers. For example, the federated distillation fusion (FedDF) (Lin T et al., 2020) improves model aggregation efficiency by using the aggregation model in FedAvg as the student model and performing ensemble distillation to capture knowledge from all client–teacher models. Recently, some works (Fallah et al., 2020; Lee et al., 2022) have focused on personalized FL (PFL). FedPHP (Li XC et al., 2021) advocates that hard-won personalized models should be rationally exploited, and proposes the concept of "inHerited Private Model" to retain the historical valuable personalization knowledge in FL. Additionally, there has been a lot of discussion in the academic community on how non-IID data may affect the structure of deep neural networks. Luo et al. (2021) found that the classifier layer in the classifier calibration with virtual representation (CCVR) deviates more than the other layers by analyzing the impact of non-IID data on each layer of a deep classification model. By re-training the classifier with virtual features taken from an approximated Gaussian mixture model, CCVR improves the performance of the global model. Federated learning with restricted softmax (FedRS) (Li XC and Zhan, 2021) states that the softmax classification layer is more susceptible to non-IID. FedRS uses restricted softmax in local processes to provide more precise updates. Although these techniques somewhat address the issue of non-IID data, they still fall short when the distribution to each client is of the long-tailed type.

2.2 Long-tailed learning

According to the research of Chen ZH et al. (2022), we divide long-tailed learning into centralized long-tailed learning and federated long-tailed learning.

2.2.1 Centralized long-tailed learning

The two primary approaches used in early centralized long-tailed learning are re-sampling and re-weighting. The data re-sampling strategy performs balanced sampling of the training data to address the imbalance of cross-class data, such as

undersampling (Pouyanfar et al., 2018) for the head classes and oversampling (Han et al., 2005) for the tail classes. The latest research by Bai et al. (2023) used out-of-distribution data in the contrastive with out-of-distribution data for long-tail learning (COLT) to dynamically re-balance the feature space to address the problem of long-tailed distribution in self-supervised learning. Re-weighting techniques (Zhang et al., 2021; Alshammari et al., 2022) give each class, or even training sample, a variable weight with the goal of modifying their gradients. Recent approaches decouple the representation learning and classifier learning phases (Kang et al., 2020) rather than training them simultaneously. For example, Zhou et al. (2020) proposed a two-branch Siamese model, named bilateral-branch network (BBN), in which the weights between features from the instance-balanced sampling branch and the reversed sampling branch were dynamically allocated. However, since the global class distribution is required by the majority of methods, this behavior violates the privacy rules in FL. Therefore, a large number of long-tailed learning approaches are not appropriate for FL scenarios.

2.2.2 Federated long-tailed learning

Recently, due to the widespread use of FL in the machine learning field, several studies (Wang LX et al., 2021; Shen YH et al., 2023) focused on the class imbalance problem in FL. For example, Shen ZB et al. (2022) proposed an agnostic constrained learning formulation, named class imbalance federated learning (CLIMB), to achieve client-level re-weighting by providing more weights on the server for clients with large local training losses. Fed-Focal Loss (Sarkar et al., 2020) directly uses focal loss (Lin TY et al., 2017) in local training, which is more focused on sample difficulty. The latest discovery is that heterogeneous data in FL cause learning difficulties due to inconsistent sample's posterior probabilities. By using probability-corrected softmax instead of softmax in the training process, Lan et al. (2022) proposed probabilistic-correction losses to align the outputs of different local models and accelerate convergence using pre-determined prototypes. However, long-tailed distribution is an extreme class imbalance with severely missing tail classes. Therefore, the above methods may perform poorly in the presence of both non-IID and long-tailed data. Most recently, the classifier re-training with federated features (CReFF) (Shang et al., 2022a) suggests that the classifier is re-trained using a set of pseudo features generated on the server. Based on CReFF, Yang et al. (2023) further improved the accuracy by integrating the real data of clients with global gradient prototypes. Zeng et al. (2023) analyzed a global objective gap between global and local re-balance strategies in the label-distribution-agnostic ensemble (LDAE). Nevertheless, they failed to recognize the significance of representation learning during the training process. The work most similar to ours is the federated ensemble distillation with imbalance calibration (FEDIC) (Shang et al., 2022b). In contrast to FEDIC, we increase the prediction accuracy by decoupling the global model and the knowledge without additional unlabeled datasets.

3 Preliminaries

3.1 Problem definition and notation

Under the non-IID and long-tailed distribution of FL setting, we consider a typical FL system with K clients and one server. K clients are associated with K datasets $D^1, D^2, ..., D^K$, where D^k denotes that the dataset of client k is of the non-IID type and accessible only by the respective client. Formally, the k^{th} client has a local dataset $D^k = \{(x_i^k, y_i^k)\}_{i=1}^{N_k}$ with N_k samples, where $y_i^k \in \{1, 2, ..., C\}$ and C is the number of classes. Without having access to any local data, our goal is to learn a global model on the server over the union of all these datasets $D \triangleq \bigcup_{k \le K} D^k$. However, in our scenario, D is drawn from a long-tailed distribution; n_c^k is defined as the number of samples of class c on client k, and $n_c = \sum_{c=1}^{C} n_c^k$. Referring to the common assumption of general long-tailed learning (Shang et al., 2022a), descending cardinal order is used to arrange the classes, i.e., if $c_1 < c_2$, then $n_{c1} \ge n_{c2}$. In addition, in the long-tailed setting, we must ensure that the header class is much larger than the tail class, i.e., $n_1 \gg n_C$. In general, a deep neural network $\phi_{\boldsymbol{w}}$ with the parameter $\boldsymbol{w}$ works as the model in FL. Specifically, the deep neural network $\phi_{\boldsymbol{w}}$ usually contains the feature extractor $f_{\boldsymbol{w}}$ and the classifier $h_{\boldsymbol{w}}$. Generally, the last classification layer in the neural network is selected as the classifier, and its output logits represent the confidence score for each class.

Furthermore, the k^{th} client model is denoted as $\phi_{\boldsymbol{w}_k}$ with parameters $\boldsymbol{w}_k$. In short, we need to compute a set of optimal model parameters $\boldsymbol{w}_{\text{g}}$.

3.2 FedAvg algorithm

FedAvg (McMahan et al., 2017) is the basic algorithm of FL. For each communication round t, the global model $\boldsymbol{w}_{\text{g}}$ is sent to the clients by the server. The clients load the received global model $\boldsymbol{w}_k = \boldsymbol{w}_{\text{g}}$. Then, the clients update their local model $\boldsymbol{w}_k$ with local dataset $D^k,\ k \in \{1, 2, \cdots, K\}$:

$$\boldsymbol{w}_k^t \leftarrow \boldsymbol{w}_k^{t-1} - \eta \nabla_{\boldsymbol{w}_k^{t-1}} L(\boldsymbol{w}_k^{t-1}; D^k), \tag{1}$$

where η is the learning rate and L is the cross-entropy loss. Subsequently, the server randomly selects m ($m < K$) clients to upload their updated models. In this round, we define the set of selected client models as S^t. The server updates the global model by using the weighted-average method:

$$\boldsymbol{w}_{\text{g}}^t = \sum_{k \in S^t} \frac{\left|D^k\right|}{\sum\limits_{k \in S^t} \left|D^k\right|} \boldsymbol{w}_k^t. \tag{2}$$

4 Propose method

Despite the fact that FedAvg assists the distributed training paradigm, the overall accuracy is severely degraded in FL for non-IID and long-tailed distribution. In addition, compared to the head class, the accuracy of the tail class is substantially lower. So, we propose the decoupled FL to decouple the global model into the feature extractor and the classifier. We specifically suggest a client confidence re-weighting method to calibrate the biased feature extractor. At the same time, in the other perspective, we use a cRB scheme to adjust the biased classifier. However, the generalization performance of the aggregated global model on the tail class is poor when the number of tail classes is small. Therefore, we propose an ensemble calibration method to weigh the output of the two perspectives. Finally, we use logits-based ensemble distillation to distill unbiased knowledge from the calibrated ensemble model to the global model. Furthermore, due to the varying degrees of imbalance between the server and the clients, we follow the recommendation of Jiang et al. (2021b) to use a modest public dataset $D^0 = \left\{\left(x_i^0,\ y_i^0\right)\right\}_{i=1}^{N_0}$ on the server. Note that there is no data sharing or transmission in our problem scenario and that all data are separately acquired by the server. The architecture of FedDDC is shown in Fig. 1.

4.1 Decoupled FL

4.1.1 TLCR method

The local training model will deviate much from the global goal because of the difference in the distribution of local and global data. The global model generated by direct aggregation may become inaccurate, resulting in poor aggregation performance (Li XC and Zhan, 2021). Hence, we propose the TLCR method to reduce the detrimental effects of non-IID and long-tailed distribution and calibrate the biased feature extractor during the aggregation phase. TLCR personalizes the weighting for each client to minimize the influence of biased models. To understand the performance of each client on each class and obtain the consistency goal, we calculate the cross-entropy loss $L_{\text{CE}}(\boldsymbol{w}_k^t; D^0)$ between the prediction output of the local model $\boldsymbol{w}_k$ on D^0 and the ground-truth label. A lower cross-entropy loss $L_{\text{CE}}(\boldsymbol{w}_k^t; D^0)$ indicates a more confident client k about how well the data sample x^0 performs, and vice versa. Thus, we give more weights to low-loss client models than to high-loss client models. Formally, we set the loss confidence on the public dataset D^0 for each selected client, expressed as follows:

$$P_k^t(D^0) = \frac{1}{L_{\text{CE}}^{k,t}(\boldsymbol{w}_k^t; D^0)}, \tag{3}$$

where $L_{\text{CE}}^{k,t}(\boldsymbol{w}_k^t; D^0)$ denotes the loss function that is minimized in this step. The cross-entropy loss $L_{\text{CE}}^{k,t}(\boldsymbol{w}_k^t; D^0)$ in the t^{th} communication round at the k^{th} client on D^0 is calculated as follows:

$$L_{\text{CE}}^{k,t}(\boldsymbol{w}_k^t; D^0) = -\sum_{i=1}^{N_0} \sum_{c=1}^{C} I(y_i = c) \log(p_{i,c}^{k,t}), \tag{4}$$

where $I(\cdot)$ is the indication function. The predicted distribution of scores for each class is defined by the softmax function:

$$p_{i,c}^{k,t} = \frac{\exp(\boldsymbol{z}_{i,c}^{k,t})}{\sum_{j=1}^{C} \exp(\boldsymbol{z}_{i,j}^{k,t})}. \tag{5}$$

We formulaically express the prediction model $\phi_{\boldsymbol{w}_k^t}$ of the k^{th} client as $F(\cdot|\boldsymbol{w}_k^t)$. Therefore, the

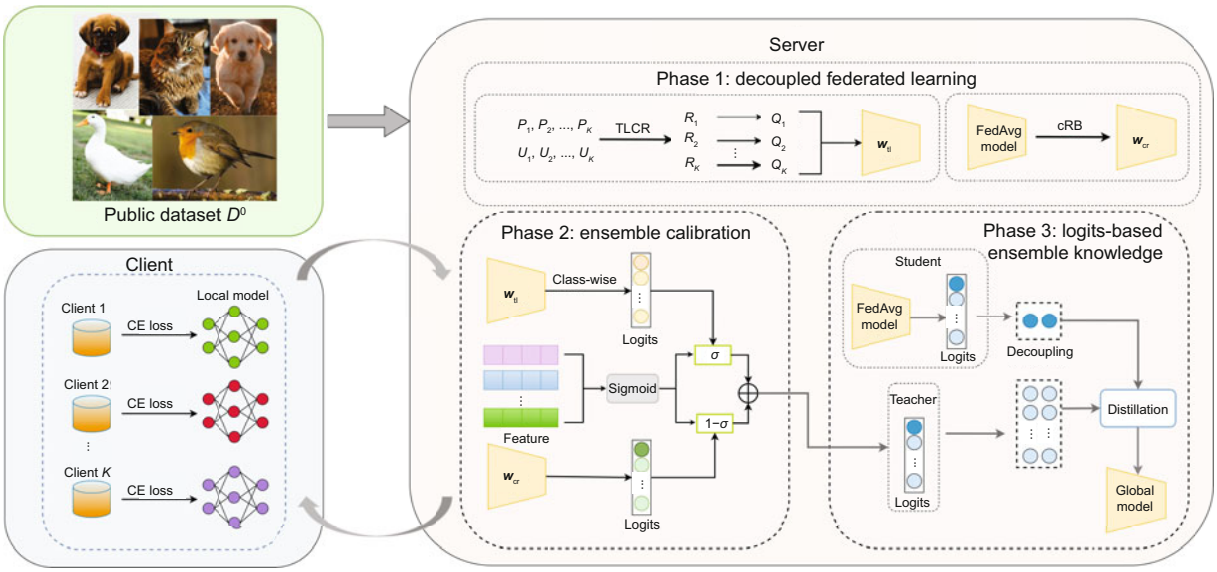

Fig. 1 Framework of FedDDC

In each round, clients send updated local models to the server, and the server sends the aggregated global model to the clients. In the decoupled federated learning, target-loss confidence re-weighting (TLCR) and classifier re-balancing (cRB) are adopted to calibrate the biased feature extractor and the biased classifier, respectively. Here, $\boldsymbol{w}_{\mathrm{tl}}$ is the model with the unbiased feature extractor. $\boldsymbol{w}_{\mathrm{cr}}$ indicates the model with the unbiased classifier

scores $\boldsymbol{z}_i^{k,t}$ of each output class from each model are defined as follows:

$$\boldsymbol{z}_i^{k,t} = F(x_i^0|\boldsymbol{w}_k^t). \tag{6}$$

To maximally quantify the confidence of each client on different classes, we calculate the logits of each sample corresponding to the target class. Specifically, the target confidence of each client on each sample is represented by the sum of the sample target logits:

$$U_k^t\left(D^0\right) = \sum_{i=1}^{N_0} \sum_{c=y_i} \boldsymbol{z}_{i,c}^{k,t}, \tag{7}$$

where $\boldsymbol{z}_{i,c}^{k,t}$ indicates the predicted value of the c^{th} class of the k^{th} client on the public dataset D^0 in round t.

The goal of target confidence is to obtain the logit that matches the ground-truth label class. By considering both loss and target confidence, the k^{th} client confidence in round t can be defined as

$$R_k^t = P_k^t\left(D^0\right) U_k^t\left(D^0\right). \tag{8}$$

It measures the confidence for each client separately by quantifying the loss magnitude and the target score on the public dataset. Then, we re-weight the clients according to their level of confidence so that the global model can learn more from confident clients and less from unconfident clients. The weight given to the k^{th} client in round t is determined by the client confidence Q_k^t, which is demonstrated as follows:

$$Q_k^t = \frac{\exp(R_k^t)}{\sum_{k=1}^K \exp(R_k^t)}. \tag{9}$$

The above weighted regularization can minimize the influence of non-IID and long-tailed distribution. With the help of TLCR, we dynamically assign weights to the model learned by the client in each round, resulting in an aggregated model with the unbiased feature extractors, as follows:

$$\boldsymbol{w}_{\mathrm{tl}}^t = \sum_{k=1}^K Q_k^t \boldsymbol{w}_k^t. \tag{10}$$

4.1.2 cRB scheme

A large number of studies (Li XC and Zhan, 2021; Luo et al., 2021) show that the biased classifier has the greatest influence on deep neural networks. With the idea of decoupling learning (Kang et al., 2020), we use the method of cRB to obtain the unbiased classifier. Thus, we suggest retraining the biased classifier. Specifically, we re-balance the

FedAvg model $\boldsymbol{w}_{\mathrm{g}}^t$ on the public dataset D^0 to produce a model with an unbiased classifier:

$$\boldsymbol{w}_{\mathrm{cr}}^t \leftarrow \boldsymbol{w}_{\mathrm{g}}^t - \eta \nabla_{\boldsymbol{w}_{\mathrm{g}}^t} L(\boldsymbol{w}_{\mathrm{g}}^t; D^0). \tag{11}$$

Since D^0 is balanced, $\boldsymbol{w}_{\mathrm{cr}}$ obtains an unbiased classifier through re-training.

With the help of TLCR and cRB, we obtain the unbiased feature extractor and the unbiased classifier. However, the above method still does not work well when the long-tailed distribution has a fairly small number of tail classes.

4.2 Ensemble calibration

We propose an ensemble calibration approach based on logit and feature to mitigate the negative effects of a very limited number of tail classes. The study of Zhao BR et al. (2022) shows that logits have a higher semantic level than deep features. Thus, for convenience, the two models derived from decoupled FL need to be adjusted to the logit form.

In the above method, while confidence based on TLCR can yield a more reliable consensus from the clients, data imbalances across the participating models are less resolved. We use the class-wise logit adjustment (Shang et al., 2022b) to further enhance the logits of the tail classes by learnable parameters $\boldsymbol{a}_z, \boldsymbol{b}_z \in \mathbb{R}^C$. Thus, the calibrated logits $\boldsymbol{z}_{\mathrm{tl}}^t$ after linear transformation on each class are defined as follows:

$$\boldsymbol{z}_{\mathrm{tl}}^t = \boldsymbol{a}_z \cdot F(x_i^0 | \boldsymbol{w}_{\mathrm{tl}}^t) + \boldsymbol{b}_z. \tag{12}$$

The logits of the re-balanced model may then be calculated using the following formula:

$$\boldsymbol{z}_{\mathrm{cr}}^t = F(x_i^0 | \boldsymbol{w}_{\mathrm{cr}}^t). \tag{13}$$

The logits $\boldsymbol{z}_{\mathrm{tl}}^t$ and $\boldsymbol{z}_{\mathrm{cr}}^t$ have been adjusted to take into account the non-IID and long-tailed distribution coming from different sources. That is, $\boldsymbol{z}_{\mathrm{tl}}^t$ are generated based on confidence re-weighting and have an unbiased feature extractor, while $\boldsymbol{z}_{\mathrm{cr}}^t$ are based on cRB and have an unbiased classifier.

Although these two models can categorize objects accurately to a certain extent, the prediction results of the ensemble model are more reliable than those of a single model (Lin T et al., 2020). Thus, inspired by Zhang et al. (2021), we propose an adaptive calibration function based on logits and features to control the trade-off between $\boldsymbol{z}_{\mathrm{tl}}^t$ and $\boldsymbol{z}_{\mathrm{cr}}^t$, so as to achieve ensemble calibration. Specifically, we define a calibration function $\sigma(\boldsymbol{x})$ to adaptively combine $\boldsymbol{z}_{\mathrm{tl}}^t$ and $\boldsymbol{z}_{\mathrm{cr}}^t$:

$$\boldsymbol{z}_{\mathrm{ec}}^t = \sigma(\boldsymbol{x}) \boldsymbol{z}_{\mathrm{tl}}^t + (1 - \sigma(\boldsymbol{x})) \boldsymbol{z}_{\mathrm{cr}}^t, \tag{14}$$

where $\boldsymbol{x}$ is $x_i^0 \in D^0$ in our setting. The calibration function $\sigma(\boldsymbol{x})$ is formulated as follows:

$$\sigma(\boldsymbol{x}) = \mathrm{sigmoid}(\boldsymbol{v}^{\mathrm{T}} \boldsymbol{g}(\boldsymbol{x})). \tag{15}$$

Here $\boldsymbol{g}(\boldsymbol{x})$ is composed of the feature ensemble and the re-trained feature, and $\boldsymbol{v} \in \mathbb{R}^d$ is a learnable parameter. Thus, $\boldsymbol{g}(\boldsymbol{x})$ can be expressed as follows:

$$\boldsymbol{g}(\boldsymbol{x}) = \frac{1}{|S^t|} \sum_{k \in S^t} f_{\boldsymbol{w}_k^t}(\boldsymbol{x}) f_{\boldsymbol{w}_{\mathrm{cr}}^t}(\boldsymbol{x}), \tag{16}$$

where $|S^t|$ stands for the total number of clients selected in each round.

4.3 Logits-based ensemble distillation

Although the ensemble model with better performance is obtained with the help of global ensemble calibration, the global model does not get any knowledge from the calibrated ensemble model. Therefore, we suggest using KD (Hinton et al., 2015) to distill knowledge from the teacher model (i.e., the calibrated ensemble model) to the student model (i.e., the global model). However, the essence of the calibrated ensemble model is the ensemble logits obtained by calibration and adjustment based on the class scores of the model output, which limits most methods in KD. To extract knowledge from the teacher model to the student model as fully and efficiently as possible, we propose logits-based ensemble distillation inspired by decoupled KD (DKD) (Zhao BR et al., 2022). As far as we know, this is the first application of DKD to non-IID and long-tailed distribution in FL. We re-formulate the classical KD loss into two parts. Specifically, we define the binary probabilities $\boldsymbol{b}^t = [p_c^t, p_{\setminus c}^t] \in \mathbb{R}^C$ of the target class (p_c^t) and all other non-target classes ($p_{\setminus c}^t$) to segregate the predictions relevant and irrelevant to the target class:

$$p_c^t = \frac{\exp(\boldsymbol{z}_c^t)}{\sum_{j=1}^C \boldsymbol{z}_j^t}, \quad p_{\setminus c}^t = \frac{\sum_{k=1, k \neq c}^C \exp(\boldsymbol{z}_k^t)}{\sum_{j=1}^C \boldsymbol{z}_j^t}. \tag{17}$$

We define $\hat{\boldsymbol{p}}^t = [\hat{p}_1^t, \hat{p}_2^t, ..., \hat{p}_{c-1}^t, \hat{p}_{c+1}^t, \hat{p}_{c+2}^t, ..., \hat{p}_C^t] \in \mathbb{R}^{C-1}$ to model the probabilities without considering the c^{th} class:

$$\hat{p}_i^t = \frac{\exp(\boldsymbol{z}_i^t)}{\sum_{j=1, j \neq c}^C \exp(\boldsymbol{z}_j^t)}. \tag{18}$$

The traditional KD using Kullback–Leibler (KL) divergence (i.e., $\mathrm{KL}(p_{\mathrm{T}}^t||p_{\mathrm{S}}^t)$, where T and S represent teacher and student, respectively) as a loss function is described by the following expression:

$$\mathrm{KD}=p_{\mathrm{T},c}^t\log\left(\frac{p_{\mathrm{T},c}^t}{p_{\mathrm{S},c}^t}\right)+\sum_{i=1,i\neq c}^{C}p_{\mathrm{T},i}^t\log\left(\frac{p_{\mathrm{T},i}^t}{p_{\mathrm{S},i}^t}\right). \tag{19}$$

According to Eqs. (18) and (19), we have $\hat{p}_i^t=p_c^t/p_{\setminus c}^t$; so, we can rewrite Eq. (19) as follows:

$$\begin{aligned}\mathrm{KD}=&\underbrace{p_{\mathrm{T},c}^t\log\left(\frac{p_{\mathrm{T},c}^t}{p_{\mathrm{S},c}^t}\right)+p_{\mathrm{T},\setminus c}^t\log\left(\frac{p_{\mathrm{T},\setminus c}^t}{p_{\mathrm{S},\setminus c}^t}\right)}_{\mathrm{KL}(b_{\mathrm{T}}^t||b_{\mathrm{S}}^t)}\\&+\underbrace{p_{\mathrm{T},\setminus c}^t\sum_{i=1,i\neq c}^{C}\hat{p}_{\mathrm{T},i}^t\log\left(\frac{\hat{p}_{\mathrm{T},i}^t}{\hat{p}_{\mathrm{S},i}^t}\right)}_{\mathrm{KL}(\hat{p}_{\mathrm{T}}^t||\hat{p}_{\mathrm{S}}^t)}\\&=\mathrm{KL}(b_{\mathrm{T}}^t||b_{\mathrm{S}}^t)+(1-p_{\mathrm{T},c}^t)\mathrm{KL}(\hat{p}_{\mathrm{T}}^t||\hat{p}_{\mathrm{S}}^t),\end{aligned} \tag{20}$$

where $\mathrm{KL}(b_{\mathrm{T}}^t||b_{\mathrm{S}}^t)$ refers to the degree of similarity between teacher's and student's binary probabilities of the target class, named target class KD (TCKD). $\mathrm{KL}(\hat{p}_{\mathrm{T}}^t||\hat{p}_{\mathrm{S}}^t)$ stands for the degree of similarity between teacher's and student's probabilities among non-target classes, named non-target class KD (NCKD). According to Zhao BR et al. (2022), the expression for KD is rewritten as DKD. Specifically, we introduce two hyperparameters α and β as the weights for TCKD and NCKD, respectively. The following is a definition of the DKD loss function:

$$\mathrm{DKD}=\alpha\mathrm{TCKD}+\beta\mathrm{NCKD}. \tag{21}$$

According to Hinton et al. (2015), we construct a KD formula with two loss components: (1) L_{CE} is the cross-entropy loss between the global model and the ground-truth label; (2) L_{DKD} is the DKD loss between the calibrated ensemble model and the global model. Thus, the global loss with hyperparameter $\lambda\in[0,1]$ can be defined as follows:

$$L_{\mathrm{Global}}=\lambda L_{\mathrm{CE}}+(1-\lambda)L_{\mathrm{DKD}}. \tag{22}$$

Finally, the global model extracts knowledge based on the ensemble logits $\boldsymbol{z}_{\mathrm{ec}}^t$ and the public dataset D^0. Specifically, the model parameters are updated as follows:

$$\boldsymbol{w}_{\mathrm{g}}^t\leftarrow\boldsymbol{w}_{\mathrm{g}}^t-\eta\nabla_{\boldsymbol{w}_{\mathrm{g}}^t}L_{\mathrm{Global}}(\boldsymbol{z}_{\mathrm{ec}}^t|\boldsymbol{w}_{\mathrm{g}}^t;D^0). \tag{23}$$

These procedures are iterated for a finite number of rounds. The overall procedures are summarized in Algorithm 1.

Algorithm 1 Training process of FedDDC

Input: Initialized global model $\boldsymbol{w}^0$, public dataset D^0, number of steps I for decoupled federated learning, number of steps J for logits-based ensemble distillation, number of communication rounds T
Output: Global model $\boldsymbol{w}_{\mathrm{g}}^T$ on round T

for $t=1$ to T **do**
  Randomly select a set of active clients S^t
  for $k\in S^t$ **do**
    Update local model $\boldsymbol{w}_k^t$ using Eq. (1)
    Send $\boldsymbol{w}_k^t$ to the server
  end for
  Aggregate local models to $\boldsymbol{w}_{\mathrm{g}}^t$ using Eq. (2)
  for $i=1$ to I **do**
    Aggregate re-weighting models to $\boldsymbol{w}_{\mathrm{tl}}^t$ using Eq. (10)
    Re-balance the model of $\boldsymbol{w}_{\mathrm{g}}^t$ to $\boldsymbol{w}_{\mathrm{cr}}^t$ using Eq. (11)
  end for
  Compute decoupled federated learning outputs $\boldsymbol{z}_{\mathrm{tl}}^t$ and $\boldsymbol{z}_{\mathrm{cr}}^t$ using Eqs. (12) and (13), respectively
  Compute ensemble logits $\boldsymbol{z}_{\mathrm{ec}}^t$ using Eq. (14)
  for $j=1$ to J **do**
    Extract ensemble knowledge $\boldsymbol{z}_{\mathrm{ec}}^t$ to the global model $\boldsymbol{w}_{\mathrm{g}}^t$ using Eq. (23)
  end for
  Send $\boldsymbol{w}_{\mathrm{g}}^t$ to clients
end for

5 Experiments

5.1 Experimental settings

5.1.1 Datasets

We validate our method with three widely used image classification benchmarks: MNIST (Lecun et al., 1998), CIFAR-10, and CIFAR-100 (Krizhevsky, 2009). MNIST is a 10-class dataset for image classification of handwritten digits. The data samples in MNIST are 28×28 single-channel images. CIFAR-10 consists of 60 000 32×32 color images in 10 classes, with 6000 images per class. CIFAR-100 consists of 60 000 32×32 color images in 100 classes, with 600 images per class. Note that the 100 classes in CIFAR-100 are grouped into 20 superclasses. The three datasets have different recognition difficulties: MNIST is the least semantically complicated and is

relatively straightforward; CIFAR-10 is more challenging as it has more semantic gaps; CIFAR-100 has the most complex composition and is the most difficult to recognize. In the following, we show how to simulate non-IID and long-tailed versions of these three datasets in FL.

5.1.2 Non-IID data

To simulate a non-IID scenario in FL, we allocate samples to local clients according to the Dirichlet distribution (Yurochkin et al., 2019; Lan et al., 2022). Specifically, we sample $p_c \sim \mathrm{Dir}_K(\alpha)$ and assign a $p_{c,k}$ proportion of the samples in class c to client k. $\mathrm{Dir}(\cdot)$ denotes the Dirichlet distribution. The value of α controls the degree of non-IID-ness. The smaller the α, the higher the degree of non-IID-ness. Fig. 2 shows how data samples with different values of α are dispersed among 20 clients on CIFAR-10. We set $\alpha = 0.1$ in our main experiments, and present the results of $\alpha \in \{0.1, 1, 10\}$ in the ablation study in Section 5.3.3.

5.1.3 Long-tailed data

In real-world situations, there is no guarantee that the classes of global data will follow a uniform distribution. Therefore, we follow existing studies (Cao et al., 2019; Cui et al., 2019; Shang et al., 2022a) to build the long-tailed versions of training sets for the above three datasets: MNIST-LT, CIFAR-10-LT, and CIFAR-100-LT. Specifically, we define the term imbalance factor (IF) as $\mathrm{IF} = \frac{\max_c\{n_c\}}{\min_c\{n_c\}}$, which denotes the ratio between the sample sizes of the head class and the tail class, to measure the degree of long-tailed distribution. The larger the value of IF, the higher the degree of long-tailed distribution. The number of training samples for each class n_c follows an exponential function $n_c = n \times \mathrm{IF}^{c/(C-1)}$, where n is the number of training samples for the first class.

Fig. 3 shows the number of training images per class on CIFAR-10-LT with IF$\in$ {10, 20, 50, 100, 200}. With the increase of IF, the number of head classes changes relatively little, while the number of tail classes sharply declines. In our main experiments, we use IF $\in \{10, 50, 100\}$ to simulate different degrees of long-tailed distribution. To further analyze the data conditions for the joint scenario of non-IID and long-tailed data in FL, we show the visualization images of IF $\in \{10, 50, 100\}$, $\alpha \in \{0.1, 10\}$, and $K = 20$ on CIFAR-10-LT in Fig. 4. It is clear that with the increase of IF and the decrease of α, the data distribution becomes more heterogeneous, with fewer samples belonging to the tail class.

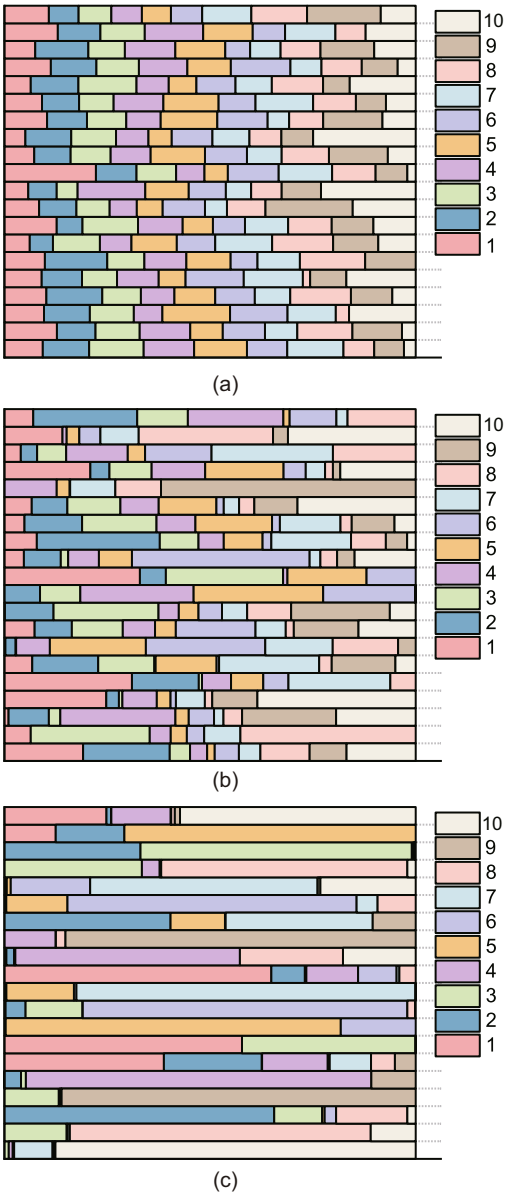

Fig. 2 Visualization of data heterogeneity with different α's among 20 clients on CIFAR-10: (a) non-IID, α=10; (b) non-IID, α=1; (c) non-IID, α=0.1 (The horizontal indicates the proportion of different classes in CIFAR-10 and the vertical indicates 20 different clients)

5.1.4 Baseline methods

We contrast FedDDC with FL methods to demonstrate the effectiveness of FedDDC on non-IID

data, including FedAvg (McMahan et al., 2017), FedAvgM (Hsu et al., 2019), FedProx (Li T et al., 2020), FedNova (Wang JY et al., 2020), and SCAFFOLD (Karimireddy et al., 2020). Then, we compare FedDDC with the federated KD algorithm FedDF (Lin T et al., 2020) and FedBE (Chen HY and Chao, 2021) under the same setting. To demonstrate the validity of FedDDC in long-tailed data, we compare it with imbalance-oriented FL methods, including Fed-Focal Loss (Sarkar et al., 2020), Ratio Loss (Wang LX et al., 2021), and FedAvg with τ-norm (Kang et al., 2020). Moreover, we compare the solution to both non-IID and long-tailed distribution in FL, including FEDIC (Shang et al., 2022b).

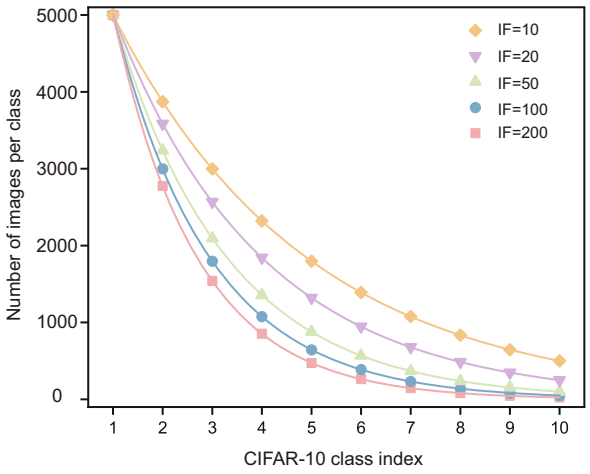

Fig. 3 Number of training samples per class in CIFAR-10-LT with different IF's

5.1.5 Training details

All experiments are run using PyTorch on NVIDIA GeForce RTX3060 graphics processing unit (GPU). By default, we run $T = 200$ global communication rounds and use standard cross-entropy loss. We fix the total number of clients at 20 and the percentage of active clients at 40% in each round. We randomly discard 5% of our clients during training to simulate a real scene. For local training, we set the batch size at 128 and use stochastic gradient descent (SGD) with a learning rate of 0.1 as the optimizer. For server training, we set the number of decoupled FL steps I at 100, the number of distillation steps J at 100, and use Adam with a learning rate of 0.002 for the optimizer.

We define the public dataset D^0 as $1/50$ of the samples for each class in the training set. According to Zhao BR et al. (2022), the hyperparameters α and β of DKD loss are fixed as 1.0 and 2.0, respectively.

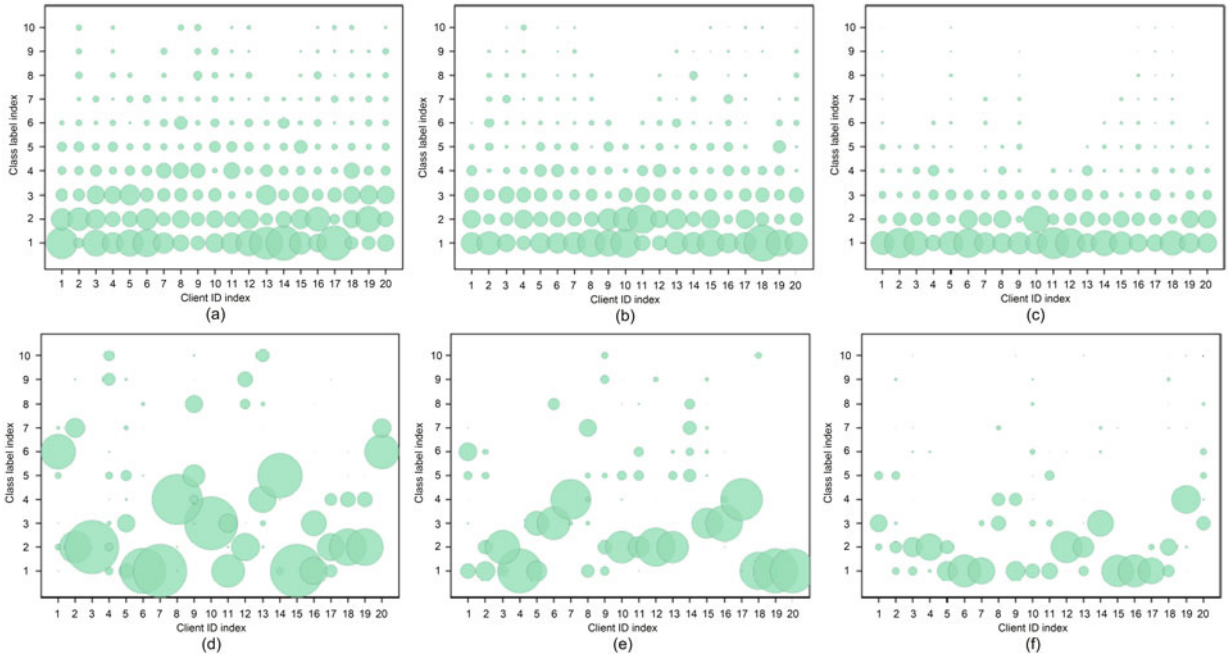

Fig. 4 Visualization of 20 clients on CIFAR-10-LT with different α's and IF's: (a) $\alpha = 10$ and IF=10; (b) $\alpha = 10$ and IF=50; (c) $\alpha = 10$ and IF=100; (d) $\alpha = 0.1$ and IF=10; (e) $\alpha = 0.1$ and IF=50; (f) $\alpha = 0.1$ and IF=100 (The size of the dot reveals how many samples are assigned to each class. The larger the dot size, the greater the number of samples)

The hyperparameter λ of our method is fixed at 0.5 in our main experiments. We explore the effect of different hyperparameters λ in the ablation study in Section 5.3.2. We use the convolutional neural network (CNN) (McMahan et al., 2017) for MNIST-LT and residual network ResNet-8 (He KM et al., 2016) for CIFAR-10/100-LT.

5.2 Main results

In Table 1, we show the best test accuracy achieved of FedDDC and the baselines for different degrees of IFs. The pivotal conclusion drawn from this table is that our technique consistently outperforms the other methods in all circumstances. As we can see, when IF=100, our method improves the baseline FedAvg accuracy by 2.37% under MNIST-LT, 25.45% under CIFAR-10-LT, and 42.08% under CIFAR-100-LT. Since FedAvgM, FedProx, SCAFFOLD, and FedNova handle only non-IID data without considering the long-tailed distribution, they are less effective than FedDDC. For example, while SCAFFOLD uses variables to correct the client drift in local updates, the aggregated global model affected by long-tailed distribution does not benefit the local model. In some settings, distillation-based FL methods are even less accurate than the baseline FedAvg, which indicates that the extracted knowledge is limited under the joint problem. For FedDF and FedBE, the knowledge of the teacher model is affected by the long-tailed distribution because the teacher model comes from each client with long-tailed data. For the imbalance-oriented FL methods, the Ratio Loss method outperforms the baseline method FedAvg. However, there is still a performance gap compared with FedDDC. The performance of Fed-Focal Loss shows that changing the weights of hard- or easy-to-classify samples in the loss function does not substantially alleviate the problem of missing tail classes. The Ratio Loss method incorporates the principle of class-level reweighting, deliberately accentuating the learning focus on tail classes. Nevertheless, it ignores the representation learning on the head classes. Compared with FEDIC, which also solves the joint problem, FedDDC further decouples the model and knowledge, improving the performance. Moreover, we use only small public datasets and do not use unlabeled datasets to aid training.

5.3 Ablation study

To further validate the performance of FedDDC, we evaluate the effect of different components, hyperparameters, and non-IID degrees. Experiments are conducted on CIFAR-10-LT in this subsection.

5.3.1 Effect of each component

To further validate the effectiveness of FedDDC, we evaluate the effect of each component, as shown in Tables 2–4. We evaluate three components in the

Table 1 Top-1 test accuracy of state-of-the-art federated learning (FL) methods on MNIST-LT, CIFAR-10-LT, and CIFAR-100-LT with different IF's

Manner	Method	Accuracy (%)								
		MNIST-LT			CIFAR-10-LT			CIFAR-100-LT		
		IF=100	50	10	100	50	10	100	50	10
FL methods for non-IID	FedAvg	93.15	95.51	98.03	51.56	52.55	58.43	24.43	27.17	37.27
	FedAvgM	93.42	95.83	97.94	53.27	55.49	60.37	23.91	27.39	38.21
	FedProx	93.47	95.24	98.00	51.99	54.33	59.10	24.67	26.70	37.88
	SCAFFOLD	93.18	95.59	98.31	51.97	53.46	59.11	24.42	27.20	38.07
	FedNova	93.51	95.79	97.91	53.89	56.26	61.01	26.42	28.37	39.29
Distillation-based FL methods	FedDF	92.25	94.42	96.88	51.31	52.30	59.23	24.82	26.55	38.34
	FedBE	91.14	93.44	96.23	48.58	49.99	52.49	21.34	24.41	30.13
Imbalance-oriented FL methods	Fed-Focal Loss	93.81	96.32	97.90	51.28	53.57	60.03	24.30	26.55	36.46
	Ratio Loss	93.69	96.07	98.07	53.24	54.17	59.71	27.01	29.57	39.74
	FedAvg+τ-norm	90.43	93.33	96.38	45.28	45.99	49.20	20.91	22.88	33.41
FL methods for non-IID and long-tailed distribution	FEDIC	94.58	96.84	98.22	63.05	64.21	66.27	33.68	35.12	42.13
	FedDDC	**95.36**	**97.10**	**98.41**	**64.68**	**65.32**	**67.78**	**34.71**	**36.21**	**45.34**

Values in bold are the best results

study: decoupled FL (DFL), ensemble calibration (EC), and logits-based ensemble distillation (LED). Note that when EC is not used, we sum the two output logits of the DFL. In the upper part (a–d) of Table 2, we analyze just the accuracy of the model without LED. As evidenced by Table 2, it becomes apparent that the remaining constituents, denoted as (b–g), exhibit enhanced accuracy relative to the benchmark algorithm designated as (a). When IF = 100, the calibrated ensemble model (d) outperforms baseline algorithm (a) by 24.56% in terms of the accuracy. Furthermore, there is only a 0.60% difference in the accuracy between the ensemble model (d) and the distillation model (g) when IF=100, which indicates that the knowledge of the ensemble model is successfully transferred. In DFL, we evaluate TLCR and cRB, separately. According to Tables 2 and 3, we observe that both TLCR (h) and cRB (i) are more effective than the baseline algorithm (a). Moreover, to demonstrate the effectiveness of LED, the traditional KD (Traditional-KD) method is used for comparison. In Table 4, we see intuitively that LED definitely performs better than Traditional-KD.

Table 2 Ablation study on the components in FedDDC on CIFAR-10-LT

Method	Component			Accuracy (%)		
	DFL	EC	LED	IF=10	50	100
(a)	–	–	–	59.71	53.66	52.24
(b)	✓	–	–	65.42	63.87	63.12
(c)	–	✓	–	63.10	61.11	60.74
(d)	✓	✓	–	**68.03**	**66.21**	**65.07**
(e)	✓	–	✓	64.85	63.14	62.40
(f)	–	✓	✓	63.76	61.67	61.15
(g)	✓	✓	✓	**67.78**	**65.32**	**64.68**

Values in bold are the best results

Table 3 Ablation study on the components in DFL on CIFAR-10-LT

Method	Component		Accuracy (%)		
	TLCR	cRB	IF=10	50	100
(h)	✓	–	63.87	63.07	62.73
(i)	–	✓	63.14	62.21	61.95

Table 4 Ablation study on the knowledge distillation in FedDDC on CIFAR-10-LT

Method	Accuracy (%)		
	IF=10	50	100
Traditional-KD	66.87	64.12	63.83
LED	**67.78**	**65.32**	**64.68**

Values in bold are the better results

5.3.2 Comparison of different hyperparameters

We conduct experiments to investigate the effect of various λ's on distillation; λ regulates the proportion of distillation loss to cross-entropy loss in Eq. (23). In Fig. 5, we can see that λ works best when it is around 0.5.

5.3.3 Influence of the degree of non-IID-ness

In our main experiments, to focus on exploring the consequences of various imbalance degrees, we fix the non-IID degree $\alpha = 0.1$. Here, we conduct extra experiments on CIFAR-10-LT with α values of 1 and 10, while fixing IF = 100. Except for the value of α, all other conditions are identical to those in the main experiments. It can be observed from Table 5 that our technique can perform at its peak under different degrees of non-IID-ness.

6 Conclusions

In this paper, we propose a dual-decoupling FL framework, FedDDC, for tackling the problem of non-IID and long-tailed distribution. FedDDC

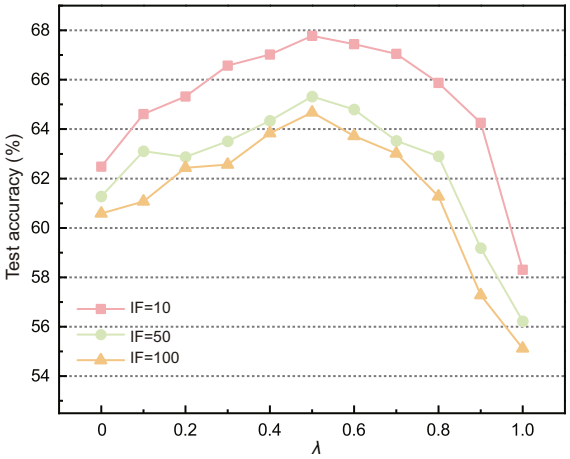

Fig. 5 Test accuracy of FedDDC with different values of λ on CIFAR-10-LT

Table 5 Different degrees of non-IID-ness on CIFAR-10-LT with IF=100

Method	Accuracy (%)		
	α=0.1	1	10
FedAvg	51.56	52.82	55.31
Fed-Focal Loss	51.28	57.99	59.37
Ratio Loss	53.24	58.43	60.14
FEDIC	63.05	65.71	66.23
FedDDC	**64.68**	**67.25**	**68.21**

Values in bold are the best results

concentrates on decoupling the global model and knowledge while decreasing the negative effects of non-IID and long-tailed data through the class-wise logit adjustment and a calibration function. In decoupled FL, we particularly propose an innovative confidence re-weighting approach and a classifier rebalancing method to calibrate the biased feature extractor and the biased classifier, respectively. To extract unbiased knowledge from the calibrated ensemble model, the logits-based ensemble distillation method is adopted to transfer knowledge. Extensive experiments prove the superior performance of our approach to cutting-edge FL methods in the setting of non-IID and long-tailed distribution. Moreover, numerous studies verify the effectiveness of each component included in our approach.

Contributors

Zhaohui WANG designed the research. Hongjiao LI and Jinguo LI supervised the research. Zhaohui WANG implemented the experiments and drafted the paper. Renhao HU and Baojin WANG helped organize the paper. Zhaohui WANG, Hongjiao LI, and Jinguo LI revised and finalized the paper.

Conflict of interest

All the authors declare that they have no conflict of interest.

Data availability

The data that support the findings of this study are openly available in public repositories. The MNIST dataset used in this study is publicly available and can be downloaded from the MNIST website (http://yann.lecun.com/exdb/mnist/). The CIFAR-10/100 datasets used in this study are also publicly available and can be downloaded from the CIFAR website (https://www.cs.toronto.edu/ kriz/cifar.html).

References

Alshammari S, Wang YX, Ramanan D, et al., 2022. Long-tailed recognition via weight balancing. IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.6887-6897. https://doi.org/10.1109/CVPR52688.2022.00677

Bai JH, Liu ZZ, Wang HL, et al., 2023. On the effectiveness of out-of-distribution data in self-supervised long-tail learning. The Eleventh Int Conf on Learning Representations.

Cao KD, Wei CL, Gaidon A, et al., 2019. Learning imbalanced datasets with label-distribution-aware margin loss. Proc 33rd Int Conf on Neural Information Processing Systems, p.1567-1578.

Chen HY, Chao WL, 2021. FedBE: making Bayesian model ensemble applicable to federated learning. https://doi.org/10.48550/arXiv.2009.01974

Chen ZH, Liu SS, Wang HL, et al., 2022. Towards federated long-tailed learning. https://doi.org/10.48550/arXiv.2206.14988

Choudhury O, Park Y, Salonidis T, et al., 2019. Predicting adverse drug reactions on distributed health data using federated learning. American Medical Informatics Association Annual Symp Proc, p.313-322.

Cui Y, Jia ML, Lin TY, et al., 2019. Class-balanced loss based on effective number of samples. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.9260-9269. https://doi.org/10.1109/CVPR.2019.00949

Fallah A, Mokhtari A, Ozdaglar A, 2020. Personalized federated learning with theoretical guarantees: a model-agnostic meta-learning approach. Proc 34th Int Conf on Neural Information Processing Systems, p.3557-3568.

Fang XW, Ye M, 2022. Robust federated learning with noisy and heterogeneous clients. IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.10062-10071. https://doi.org/10.1109/CVPR52688.2022.00983

Han H, Wang WY, Mao BH, 2005. Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning. Int Conf on Intelligent Computing, p.878-887. https://doi.org/10.1007/11538059_91

He HB, Garcia EA, 2009. Learning from imbalanced data. *IEEE Trans Knowl Data Eng*, 21(9):1263-1284. https://doi.org/10.1109/TKDE.2008.239

He KM, Zhang XY, Ren SQ, et al., 2016. Deep residual learning for image recognition. IEEE Conf on Computer Vision and Pattern Recognition, p.770-778. https://doi.org/10.1109/CVPR.2016.90

Hinton G, Vinyals O, Dean J, 2015. Distilling the knowledge in a neural network. https://doi.org/10.48550/arXiv.1503.02531

Hsu TMH, Qi H, Brown M, 2019. Measuring the effects of non-identical data distribution for federated visual classification. https://doi.org/10.48550/arXiv.1909.06335

Huang C, Li YN, Loy CC, et al., 2020. Deep imbalanced learning for face recognition and attribute prediction. *IEEE Trans Patt Anal Mach Intell*, 42(11):2781-2794. https://doi.org/10.1109/TPAMI.2019.2914680

Jalalirad A, Scavuzzo M, Capota C, et al., 2019. A simple and efficient federated recommender system. Proc 6th IEEE/ACM Int Conf on Big Data Computing, Applications and Technologies, p.53-58. https://doi.org/10.1145/3365109.3368788

Jiang ZY, Ren Y, Lei M, et al., 2021a. FedSpeech: federated text-to-speech with continual learning. Proc 30th Int Joint Conf on Artificial Intelligence, p.3829-3835. https://doi.org/10.24963/ijcai.2021/527

Jiang ZY, Chen TL, Chen T, et al., 2021b. Improving contrastive learning on imbalanced seed data via open-world sampling. https://arxiv.org/abs/2111.01004

Kang BY, Xie SN, Rohrbach M, et al., 2020. Decoupling representation and classifier for long-tailed recognition. https://doi.org/10.48550/arXiv.1910.09217

Karimireddy SP, Kale S, Mohri M, et al., 2020. SCAFFOLD: stochastic controlled averaging for federated learning. Proc 37th Int Conf on Machine Learning, p.5132-5143.

Krizhevsky A, 2009. Learning Multiple Layers of Features from Tiny Images. Technical Report, TR-2009. University of Toronto, Toronto, Canada.

Lan L, Zhang DC, Li XC, 2022. Aligning model outputs for class imbalanced non-IID federated learning. *Mach Learn*, 113:1861-1884. https://doi.org/10.1007/s10994-022-06241-5

Lecun Y, Bottou L, Bengio Y, et al., 1998. Gradient-based learning applied to document recognition. *Proc IEEE*, 86(11):2278-2324. https://doi.org/10.1109/5.726791

Lee G, Jeong M, Shin Y, et al., 2022. Preservation of the global knowledge by not-true distillation in federated learning. https://doi.org/10.48550/arXiv.2106.03097

Li DL, Wang JP, 2019. FedMD: heterogenous federated learning via model distillation. https://doi.org/10.48550/arXiv.1910.03581

Li T, Sahu AK, Zaheer M, et al., 2020. Federated optimization in heterogeneous networks. https://doi.org/10.48550/arXiv.1812.06127

Li X, Huang KX, Yang WH, et al., 2020. On the convergence of FedAvg on non-IID data. https://doi.org/10.48550/arXiv.1907.02189

Li XC, Zhan DC, 2021. FedRS: federated learning with restricted softmax for label distribution non-IID data. Proc 27th ACM SIGKDD Conf on Knowledge Discovery & Data Mining, p.995-1005. https://doi.org/10.1145/3447548.3467254

Li XC, Zhan DC, Shao YF, et al., 2021. FedPHP: federated personalization with inherited private models. European Conf on Machine Learning and Knowledge Discovery in Databases, p.587-602. https://doi.org/10.1007/978-3-030-86486-6_36

Lin T, Kong LJ, Stich SU, et al., 2020. Ensemble distillation for robust model fusion in federated learning. Proc 34th Int Conf on Neural Information Processing Systems, p.2351-2363.

Lin TY, Goyal P, Girshick R, et al., 2017. Focal loss for dense object detection. IEEE Int Conf on Computer Vision, p.2999-3007. https://doi.org/10.1109/ICCV.2017.324

Luo M, Chen F, Hu DP, et al., 2021. No fear of heterogeneity: classifier calibration for federated learning with non-IID data. https://doi.org/10.48550/arXiv.2106.05001

McMahan B, Moore E, Ramage D, et al., 2017. Communication-efficient learning of deep networks from decentralized data. Proc 20th Int Conf on Artificial Intelligence and Statistics, p.1273-1282.

Pouyanfar S, Tao YD, Mohan A, et al., 2018. Dynamic sampling in convolutional neural networks for imbalanced data classification. IEEE Conf on Multimedia Information Processing and Retrieval, p.112-117. https://doi.org/10.1109/MIPR.2018.00027

Sarkar D, Narang A, Rai S, 2020. Fed-Focal Loss for imbalanced data classification in federated learning. https://doi.org/10.48550/arXiv.2011.06283

Shang XY, Lu Y, Huang G, et al., 2022a. Federated learning on heterogeneous and long-tailed data via classifier retraining with federated features. Proc 31st Int Joint Conf on Artificial Intelligence, p.2218-2224. https://doi.org/10.24963/ijcai.2022/308

Shang XY, Lu Y, Cheung YM, et al., 2022b. FEDIC: federated learning on non-IID and long-tailed data via calibrated distillation. IEEE Int Conf on Multimedia and Expo, p.1-6. https://doi.org/10.1109/ICME52920.2022.9860009

Shen YH, Wang HX, Lv HR, 2023. Federated learning with classifier shift for class imbalance. https://doi.org/10.48550/arXiv.2304.04972

Shen ZB, Cervino J, Hassani H, et al., 2022. An agnostic approach to federated learning with class imbalance. The 10th Int Conf on Learning Representations.

Shoham N, Avidor T, Keren A, et al., 2019. Overcoming forgetting in federated learning on non-IID data. https://doi.org/10.48550/arXiv.1910.07796

Tan B, Liu B, Zheng V, et al., 2020. A federated recommender system for online services. Proc 14th ACM Conf on Recommender Systems, p.579-581. https://doi.org/10.1145/3383313.3411528

Wang JY, Liu QH, Liang H, et al., 2020. Tackling the objective inconsistency problem in heterogeneous federated optimization. Proc 34th Int Conf on Neural Information Processing Systems, p.7611-7623.

Wang LX, Xu SC, Wang X, et al., 2021. Addressing class imbalance in federated learning. *Proc AAAI Conf Artif Intell*, 35(11):10165-10173. https://doi.org/10.1609/aaai.v35i11.17219

Yang WK, Chen DL, Zhou H, et al., 2023. Integrating local real data with global gradient prototypes for classifier re-balancing in federated long-tailed learning. https://doi.org/10.48550/arXiv.2301.10394

Yu FX, Rawat AS, Menon AK, et al., 2020. Federated learning with only positive labels. Proc 37th Int Conf on Machine Learning, p.10946-10956.

Yurochkin M, Agarwal M, Ghosh S, et al., 2019. Bayesian nonparametric federated learning of neural networks. Proc 36th Int Conf on Machine Learning, p.7252-7261.

Zeng YP, Liu L, Wu BY, et al., 2023. Label-distribution-agnostic ensemble learning on federated long-tailed data. Int Conf on Learning Representations.

Zhang SY, Li ZM, Yan SP, et al., 2021. Distribution alignment: a unified framework for long-tail visual recognition. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.2361-2370. https://doi.org/10.1109/CVPR46437.2021.00239

Zhao BR, Cui Q, Song RJ, et al., 2022. Decoupled knowledge distillation. IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.11943-11952. https://doi.org/10.1109/CVPR52688.2022.01165

Zhao Y, Li M, Lai LZ, et al., 2022. Federated learning with non-IID data. https://doi.org/10.48550/arXiv.1806.00582

Zhou BY, Cui Q, Wei XS, et al., 2020. BBN: bilateral-branch network with cumulative learning for long-tailed visual recognition. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.9716-9725. https://doi.org/10.1109/CVPR42600.2020.00974

Zhu ZD, Hong JY, Zhou JY, 2021. Data-free knowledge distillation for heterogeneous federated learning. Proc 38th Int Conf on Machine Learning, p.12878-12889.