LABEL-DISTRIBUTION-AGNOSTIC ENSEMBLE LEARNING ON FEDERATED LONG-TAILED DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated Learning (FL) is a distributed machine learning paradigm that enables devices to collaboratively train a shared model. Long-tailed distribution in nature aggravates both heterogeneity and privacy issues for real-world FL applications, *e.g.*, local clients may exhibit diverse imbalanced distributions with different imbalance ratios and minority classes. More importantly, existing re-balance strategies assume that global label distribution is available, which may conflict with the privacy requirement of FL. To this end, we propose a label-distribution-agnostic ensemble learning (LDAE) framework to re-balance the global objective with privacy protection. In particular, based on the model updates of clients, we present proxy information to guide the balanced training of multiple experts with privacy-preserving. The heterogeneity issue can be alleviated via interactions among experts. Through an empirical study, we demonstrate that (1) there is a global objective gap between global and local re-balance strategies [1] and (2) with protecting data privacy, proxy information can be used as an alternative to label distribution for existing re-balance strategies. Extensive experiments on long-tailed decentralized datasets demonstrate the effectiveness of our method, showing superior performance over previous state-of-the-arts.

1 INTRODUCTION

Federated learning (FL) aims to collaboratively learn from data dominated by a number of remote clients, and produces a highly accurate global model on a server with aggregated knowledge. It has been applied in many domains, such as medical applications Li et al. (2019); Malekzadeh et al. (2021); Liu et al. (2021) and autonomous vehicles Samarakoon et al. (2019); Pokhrel & Choi (2020). To deploy FL in practice, it is necessary to deal with the privacy constraints for disparate data sources as well as the data heterogeneity, *i.e.*, data stored on different devices is not independently and identically distributed Li et al. (2020a). Moreover, due to the fact that long-tailed class imbalance is common in the real-world data Cui et al. (2019); Liu et al. (2019), it is inevitable in the decentralized datasets, leading to a worse global model.

As shown in Figure 1, the imbalance issue is aggravated due to the data heterogeneity and privacy constraints of FL Shen et al. (2021). From the local perspective, there can be a significant divergence among the imbalanced label distributions of different clients, resulting in the heterogeneous imbalance of the decentralized data, *i.e.*, the minority class of one client can actually be a majority class of another client with different imbalance distribution. From the global perspective, one should handle the imbalance issue with privacy preservation Li et al. (2021), *i.e.*, the server should not require clients to upload label distribution information for re-balance strategies.

Several prior works attempt to tackle the imbalance problem in FL Duan et al. (2020); Wang et al. (2021); Shen et al. (2021), considering the datasets with a few classes, *e.g.*, ten classes or twenty classes. These methods suffer from significant performance drops when faced with large-scale datasets with long-tailed class distribution Liu et al. (2019); Zhang et al. (2021b). On the other hand, although previous long-tailed re-balance approaches such as data re-sampling

[1] The local re-balance strategy means that each client utilizes class-balance methods based on the local label distribution, while the global re-balance strategy applies class-balance methods using global label distribution as the class prior.

Kang et al. (2019), loss re-weighting Cui et al. (2019); Ren et al. (2020); Hong et al. (2021), and gradient tuning can alleviate the imbalanced issue in centralized data, it is unknown whether they can work well for long-tailed FL problem. To this end, we theoretically and experimentally investigate the effectiveness of existing re-balance algorithms from the global and local perspectives in the long-tailed FL setting. We demonstrate that there is a gap of objective functions between the local and global re-balance strategies under heterogeneous data distribution, while global re-balance yields the matched objective under centralized data. However, global re-balance may violate the principle of privacy protection in FL Wang et al. (2021); McMahan et al. (2017), since the global server needs to collect local label distribution of each client.

In this work, we propose a **L**abel-**D**istribution-**A**gnostic **E**nsemble learning framework (LDAE) to address challenges of privacy and heterogeneity under long-tailed distribution. Specifically, based on the local model updates uploaded by clients, we present two types of proxy information as the prior information for both re-balance methods (*e.g.*, re-weighting) and client grouping, which is agnostic about the local label distributions. To further alleviate the heterogeneous imbalance issue in FL, we introduce a multi-expert architecture to assign experts for each group of clients. Each expert focuses on one corresponding client group to mitigate the negative impact of heterogeneous data distribution. Different with previous ensembled models, our method allows the information interaction among different experts trained on different local data distributions. In conclusion, our key contributions are:

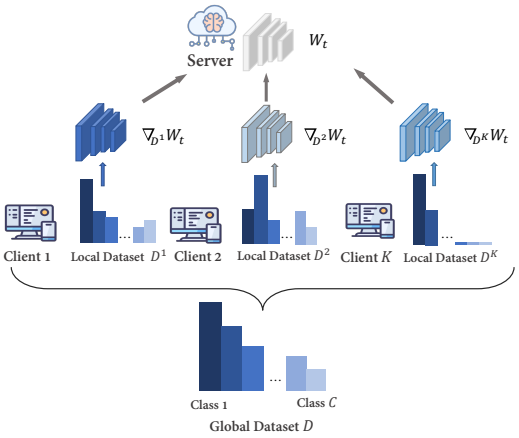


Figure 1: Illustration of heterogeneous imbalance data in FL. The local datasets of clients have different imbalanced class distributions, and their combination, i.e., the global dataset may exhibit a long-tailed class imbalance.

(1) We experimentally and theoretically demonstrate the mismatch of objectives between local and global re-balance strategies, indicating that the global re-balance strategies can well handle the long-tailed decentralized data distribution.

(2) To address imbalance issue with preserving privacy, we propose a LDAE framework called LDAE to compute two kinds of proxy information for client grouping, where heterogeneous information can be aggregated via a multi-expert architecture. Our method doesnot rely on extra privacy information (*e.g.*, local label distribution) except for the necessary updates uploaded by clients.

(3) We conduct extensive experiments on benchmark datasets. The results show that LDAE significantly outperforms state-of-the-art works, and the propose proxy information is more effective than local label distribution under the FL setting.

2 Related Work

Federated Learning. FL McMahan et al. (2017) is a learning framework proposed to protect the data privacy of participants. There is a central server and multiple clients in FL. A global model is learned on distributed data from different clients. One of the most important challenges in FL is data heterogeneity. Many previous studies concentrate on this problem Smith et al. (2017); Li et al. (2020c); Zhao et al. (2018); Sattler et al. (2020); Karimireddy et al. (2020); Li et al. (2020c); T Dinh et al. (2020). However, most of them consider that the global dataset (the dataset where all local data are centralized together) is balanced. Recently, some works Xiao & Wang (2021); Sittijuk & Tamee (2021) study the impact of class imbalance of the global dataset in FL while do not design algorithms to address it. Other works Duan et al. (2020); Wang et al. (2021); Shen et al. (2021) propose different strategies to address class imbalance in FL. However, all of their methods require extra information on server and may not be satisfied with data privacy concern. Besides, their simple re-weighting loss algorithm is not efficient enough to deal with the large-scale long-tail

problem according to our experiments. CReFF Shang et al. (2022) considers long-tailed data in FL inspired by Kang et al. (2019). However, while claiming it has no privacy concerns, their work needs feature gradients of clients' data, which can be used to recover original data with model updates.

Long-Tailed Learning. In real-world scenarios, data often has a long-tailed label distribution, where the majority classes have massive samples while the minority classes only have a few samples Zhang et al. (2021b). There are many re-balance strategies proposed from different perspectives in long-tailed learning. Class re-sampling Chawla et al. (2002); He & Garcia (2009); Kang et al. (2019) is a common type, such as over-sampling the minority classes Shen et al. (2016); Kang et al. (2019) or under-sampling He & Garcia (2009) the frequent classes. Loss re-weighting (Cost-sensitive Learning) is another scheme to learn a balanced model Sun et al. (2007); Cui et al. (2019); Lin et al. (2014). Generally, these methods tend to give a large training loss when the sample belongs to the minority class. Recently, many works in long-tailed learning focus on learning a good representation extractor to improve the generalization ability of the model. PaCo Cui et al. (2021) introduces the contrastive learning method to the long-tailed dataset. Ensemble learning based methods are also gradually playing an important role in long-tailed learning Zhou et al. (2020); Xiang et al. (2020); Wang et al. (2020); Zhang et al. (2021a); Cai et al. (2021). Until now, many re-balance strategies have made great performance gains on learning on centralized long-tailed datasets. However, it is a question of whether they can still work well in FL. In this work, we answer this question with an example and propose a novel tool to help these re-balance strategies maintain their effectiveness in FL in the following section.

3 METHOD

In this section, we highlight the challenges when long-tailed distribution encounters data heterogeneity in federated training. Then, we describe our solution, *i.e.*, GPI and multi-expert federated learning framework in detail.

Notations. We discuss a typical FL setting with K clients indexed by $[K]$ and a central server. Each client k has a local training dataset $\mathcal{D}^k$ with n^k samples, and we call the whole dataset $\mathcal{D} = \bigcup_{k\in[K]} \mathcal{D}^k$ as the global dataset. Considering a C-class classification task on the global dataset $\mathcal{D}$, each class is indexed by $[C]$ and $(\boldsymbol{x}, y) \in \mathcal{X} \times [C]$ denotes a sample in $\mathcal{D}$, where $\boldsymbol{x}$ is an image in the input space $\mathcal{X}$ and y is its corresponding label. Let n_j denote the number of training sample for class j, and $n = \sum_{j=1}^{C} n_j$ be the total number of training samples. Suppose $\mathcal{D}$ follows a long-tailed distribution, *i.e.*, the sample size is exponential distributed *w.r.t.* class index. Without loss of generality, we assume that the classes are sorted by cardinality in decreasing order, *i.e.*, if $i < j$, then $n_i \geq n_j$. The global imbalance ratio is defined as n_{max}/n_{min}. In a decentralized system, the training data on a given client is typically based on the particular data source, and hence local datasets will not be representative of the global long-tailed distribution, leading to varying imbalance ratios of local training data.

Definition 1 *(Heterogeneous Imbalance). Given a global long-tailed training dataset for FL, we define heterogeneous imbalance by the following items:(1) Different imbalanced ratios or majority (minority) classes of local datasets. (2) Different imbalanced ratios or majority (minority) classes between local and global datasets.*

The definition reflects various local imbalanced distribution *w.r.t* the global one, which is unknown due to the privacy concerns.

Typically, FL aims at learning a single shared model, and optimizing the global objective which is the aggregation of the local objective:

$$\min_{\theta\in\mathbb{R}^d} G(\theta) = \sum_{k=1}^{K} \frac{n_k}{n} F_k(\theta) \quad \text{where} \quad F_k(\theta) = \frac{1}{n_k} \sum_{(\boldsymbol{x},y)\in\mathcal{D}^k} f(\boldsymbol{x}, y, \theta) \tag{1}$$

To handle long-tailed data distribution in FL, an intuitive idea is applying a re-balance strategy on the local training to yield balanced $F_k(\theta)$ and then the server aggregates balanced model updates across clients. However, We theoretically indicate that how local re-balance affects the overall learning and

global re-balance is a better choice. Then we experimentally verify the theoretical results by the performance of the global re-balance strategy on boosting imbalanced learning.

3.1 THEORETICAL MOTIVATION

Consider a binary classification problem with either positive (+) or negative (-) ground truth. Under the imbalanced setting, we consider two clients C_0 and C_1 trained on heterogeneously imbalanced data distribution, *i.e.*, client C_0 accesses n_0^+ positives and n_0^- negatives, while client C_1 accesses n_1^+ positives and n_1^- negatives. Without loss of generality, we consider a simple loss re-weighted strategy as an example, which takes the inverse of label distribution as the samples loss weights. For the local re-balance strategy, inverse of the local label distribution is taken as the sample weights *i.e.*, n_0/n_0^+ of positive class and n_0/n_0^- of negative class for client C_0. Similarly, client C_1 uses n_1/n_1^+ for positive class and $n_1/n_1^-$for negative class. For the global re-balance strategy, inverse of the global label distribution is taken as the sample weights (*i.e.*, $(n_0+n_1)/(n_0^+ + n_1^+)$ for positive class and $(n_0+n_1)/(n_0^- + n_1^-)$ for negative class). Then we can measure the difference of global object functions between local and global re-balance.

Lemma 1 *Using the global re-balance strategy, the global objective yields the same form as the objective on the centralized dataset with re-balance strategy:*

$$G(\theta) = \frac{1}{n_0^+ + n_1^+} \sum_{(\boldsymbol{x},+)\in\mathcal{D}} f(\boldsymbol{x},+,\theta) + \frac{1}{n_0^- + n_1^-} \sum_{(\boldsymbol{x},-)\in\mathcal{D}} f(\boldsymbol{x},-,\theta) \tag{2}$$

Theorem 1 *Local Re-balance vs. Global Re-balance. Under the above setup, let Δ denotes the gap of the global objective obtained with local and global re-balance strategies, then we have:*

$$\Delta = \frac{n_1(n_0^+)^2 + n_0(n_1^+)^2}{n_0^+ n_1^+ (n_0+n_1)(n_0^+ + n_1^+)} \sum_{(\boldsymbol{x},+)\in\mathcal{D}} f(\boldsymbol{x},+,\theta) + \frac{n_1(n_0^-)^2 + n_0(n_1^-)^2}{n_0^- n_1^- (n_0+n_1)(n_0^- + n_1^-)} \sum_{(\boldsymbol{x},-)\in\mathcal{D}} f(\boldsymbol{x},-,\theta) \tag{3}$$

Interpretation. The result illustrates the following points. (1) Heterogeneous imbalance can influence the global objective via various imbalanced distributions across clients, *e.g.*, when $n_0^+/n_0^- \neq n_1^+/n_1^-$ and $n_0 \neq n_1$. (2) The local re-balance strategy will exhibit a shifted objective compared with the re-balance strategy on the global dataset. (3) The global re-balance strategy can perform better than local re-balance strategy, while the former is difficult to be obtained without violating privacy requirements of FL.

3.2 GLOBAL PROXY INFORMATION

Our theoretical findings show that global re-balance can work better than local re-balance in FL. However, due to the privacy concerns, the central server cannot ask the client to contribute its local label distribution information to obtain the global one. For most existing FL algorithms, the communications from the client to server mainly relies on gradient information. Inspired by these, we systematically explore how to take this information as the proxy for global re-balance. Suppose using cross-entropy loss to train a neural network, the gradient of the weight w^{ij} in FC layer on a sample pair (x, y) is given by:

$$\nabla_{(\boldsymbol{x},y)} \mathbf{W}_{-1}^{ij} = \frac{\partial f(\boldsymbol{x}, y, \theta)}{\partial w^{ij}}. \tag{4}$$

Previous works Zhu et al. (2019); Zhao et al. (2020) utilize the gradient to recover the label of the single image. These methods may not work in the situation where server only obtains the model parameters updates after several local epochs from clients, which is more practical in FL. From the other perspective, based on the connection between labels and gradients of the FC layer, we propose to calculate the GPI from model updates. The GPI can replace the global label distribution needed by global re-balance, resulting in a better performance than local re-balance without requiring any other information from clients. In this section, we simplify the local training process that client uses the batch gradient descend in local training, *i.e.*, the model update is the gradient on the whole local dataset. However, we use the common setting with mini-batch SGD in the experiments 4 and our proxy information still works well.

Definition 2 *(Local Proxy Information). Denote the gradient of w^{ij} after one local optimization loop on $\mathcal{D}^k$ as $\nabla_{\mathcal{D}^k}\mathbf{W}_{-1}^{ij}$ and the dimension of the input of the FC layer as H. Then the local proxy information of class i on $\mathcal{D}^k$ is defined as the gradient magnitude which is associated with i-th neuron of the FC layer:*

$$\Omega_i^k = \sum_{j=1}^{H} -\nabla_{\mathcal{D}^k}\mathbf{W}_{-1}^{ij}, \quad \textit{where} \quad \nabla_{\mathcal{D}^k}\mathbf{W}_{-1}^{ij} = \frac{\partial F_k(\theta)}{\partial w^{ij}}. \tag{5}$$

Intuitively, local proxy information (LPI) accumulates the updates of the corresponding neuron of the last FC layer over the local dataset. As indicated by the Theorem 2, LPI implicitly indicates both class frequency and sample probability over the local dataset.

Theorem 2 *For class i in a local dataset $\mathcal{D}^k$, LPI implicitly connects with its label frequency (sample number in this class) and the probability that the sample is predicted to be the current class:*

$$\Omega_i^k = \frac{1}{n^k}(n_i^k - \sum_{(x,y)\in\mathcal{D}^k} z_i) \sum_{(x,y)\in\mathcal{D}^k} \sum_{j=1}^{H} s_j \ , \tag{6}$$

where s^j is j-th input of the FC layer, y^i is i-th term of the one-hot ground truth, and z^i is the probability of being predicted as class i. We call the summation of the model output the probability term. This term is generally smaller than the label count at the early training stage when the class number C is large. It has two benefits. First, its exact value is determined by all samples in D^k, making it difficult for the server to recover the label distribution from the LPI. Second, it can be helpful to maintain the model performance on the majority class. As shown in previous researches Cui et al. (2019), the more samples in one class, the less new information the newly added samples may provide. There will be some overlapped information in the majority class and this information can be easy to learn. At the early training stage, by subtracting the probability term, we can eliminate the effect of this easily learned information at some extent. Therefore, GPI is a better representation of the proportion of majority class information than global label information. The proportion curves of the GPI are demonstrated in Figure3 to show our opinion.

Definition 3 *(Global Proxy Information). The global proxy information of class i is defined as the weighted summation of the local proxy information:*

$$\Omega_i := \sum_{k=1}^{K} \frac{n^k}{n} \mathbb{I}(\Omega_i^k > 0)\Omega_i^k \ , \tag{7}$$

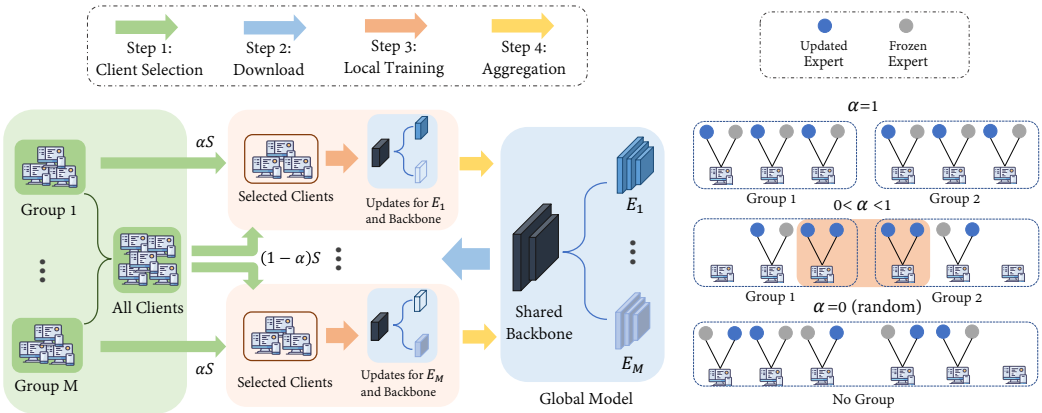
where $\mathbb{I}(\cdot)$ is the indicator function and the value is 1 if $\cdot$ is true, 0 otherwise.

When client k has few samples of class i, Ω_i^k is completely determined by the sample probability of other classes and is negative, which should be ignored. According to the above analysis, the GPI of a class is strongly related to its influence on the imbalanced training. Thus, GPI can be taken as the prior information for re-balance strategies. The performance is better than using local label distribution, as shown in the experimental section 4.

In the standard FL pipeline, only a fraction of clients are selected to conduct local updating and unloading in each training round, which poses an issue to estimate the GPI. Besides, the calculation of GPI relies on the model with initial state. Thus, our method considers all clients at the first communication round, where the global server aggregates GPI based on uploads from each client and keeps a balanced training for subsequent communication rounds.

3.3 Label-Distribution-Agnostic Ensemble Learning

Guided by our proxy information analysis, we propose a group-based ensemble learning framework to further handle heterogeneous imbalance issue for FL. Our method separates clients into different groups based on their LPI. Models trained on individual groups are ensembled together in a multi-expert framework, where each expert focuses on a group of clients with similar distributions. The



(a) Training pipeline of LDAE. (b) Client selection with different α

Figure 2: An overview of our group-based ensemble learning framework. Figure (a) shows the training pipeline with four steps in a communication round: 1) for each expert, αS clients from the corresponding group and $(1-\alpha)S$ clients from other groups are randomly selected. 2) The selected clients download current global model. 3) Each selected client updates the backbone and associated experts. 4) The server receives the updates and aggregates them into a new global model. Figure (b) different cases with different α values for client selection. When $\alpha = 0$, clients in a group only train the corresponding expert. When $0 < \alpha \leq 1$, some clients may be chosen to update multiple experts.

overall framework and the model structure are shown in Figure 2(a). The global and local models have the same structure, *i.e.*, multiple experts with a shared backbone, where each expert has individual learnable blocks.

Client Grouping. Based on the previous section 3.2, the similarity score is calculated via the cosine distance between the LPI and GPI for each client, allowing to monotonically rank the local clients. Then the ranked clients with close similarity scores are divided into the same group. Let hyper-parameter M be the number of groups, then M experts are allocated for these groups respectively, *i.e.*, group P_i corresponds to expert E_i.

Client Selection. The local datasets for expert updating come from two ways in a communication round: (1) in each group (*e.g.*, P_i), αS clients are randomly selected then corresponding expert (*e.g.*, E_i) of each selected client is updated on the local dataset, and (2) for each expert (*e.g.*, E_i), $(1-\alpha)S$ clients are randomly selected from all clients to update E_i individually. $0 \leq \alpha \leq 1$ is a hyper-parameter to control the client selection. Empirically, the difference between experts can be controlled by the value of α. As shown in Figure 2(b), if $\alpha = 1$, there is no interaction among the clients from different groups, where each group only updates the corresponding expert. If $\alpha = 0$, we adopt M times booststrap sampling from all clients, where the clients selected at i-th time are used to update E_i individually. By this way, the experts can be updated on the multiple local datasets and tends to be similar after sufficient rounds.

Expert Ensembling. After client selection step, the selected clients download current global model parameters and perform local training on the their own datasets. In particular, for each client, the classification loss is calculated based on the average logits of all experts to aggregate the knowledge of different groups, which is learned from other local datasets. Here, we adopt the balanced softmax loss (BSM) Ren et al. (2020) for local optimization due to it is simple and effective:

$$L_{\mathrm{B}} = \frac{1}{n_k} \sum_{(x,y)\in\mathcal{D}_k} -y \log \sigma \left(\frac{1}{N_e} \sum_{i=1}^{N_e} \mathbf{v}_i\left(x\right) + \log r_g^c \right), \tag{8}$$

where $\sigma(\cdot)$ is the softmax function and $\mathbf{v}_i(\cdot)$ denotes the output logits of expert. Notice that in the local training of one expert, only the backbone and corresponding expert are updated, while keeping other experts frozen. For clients selected from the client list, they may be selected multiple times to

train several experts in one round if $\alpha < 1$. During the global aggregation, the server applies all the local updates into the global model.

Table 1: Top-1 accuracy on CIFAR10-LT, CIFAR100-LT, and ImageNet-LT. The proposed GPI can be combined with existing re-balance strategies to address imbalance problem in FL.

Methods	CIFAR-10-LT		CIFAR-100-LT		ImageNet-LT
	100	50	100	50	
FedAvg	53.16	61.67	34.48	36.84	33.80
Ratio Loss	53.31	62.26	33.06	34.94	33.15
CLIMB	60.28	72.29	34.66	40.22	35.29
Focal Loss	53.88	59.00	33.88	38.03	32.85
CRT-IB	53.80	63.52	32.48	37.61	31.77
CRT-CB	63.31	69.87	34.06	39.60	35.52
LDAM + GPI	66.73	72.58	35.36	39.17	34.88
Balanced SM + GPI	67.44	74.36	37.19	41.91	37.64
LADE + GPI	64.89	71.79	38.61	41.08	38.52
RIDE + GPI	58.64	68.63	35.88	39.39	40.30
PaCo + GPI	68.84	75.35	37.76	43.24	39.04
LDAE + GPI (ours)	**71.07**	**76.85**	**40.42**	**45.16**	**45.75**

4 EXPERIMENT

Datasets. We conduct experiments on three long-tailed classification datasets: CIFAR10-LT, CIFAR100-LT Cao et al. (2019) and ImageNet-LT Liu et al. (2019). Following Cao et al. (2019), we construct CIFAR10-LT and CIFAR100-LT by an exponential decay with the controllable imbalance ratio (IR). We show the experimental results with the IR of 50 and 100.

Evaluation Protocol. We evaluate the global model on the balanced dataset using top-1 accuracy. Following typical long-tail recognition tasks, we categorize the classes into three groups: Many-shot class with > 100 samples, Medium-shot class with ≥ 20 and ≤ 100 samples and Few-shot class with < 20 samples.

Implementation Details. We use Resnet-18 He et al. (2016) for CIFAR-10-LT , Resnet-32 for CIFAR-100-LT and ResneXt as the backbone. For the FL setting, the global dataset is partitioned into 100 clients with the Dirichlet distribution to simulate the data heterogeneity Hsu et al. (2019). The degree of data heterogeneity can be controlled by α_{dir}, the parameter of Dirichlet distribution. Small α_{dir} indicates high data heterogeneity. We set α_{dir} as 0.5 for CIFAR-10-LT, 0.1 for CIFAR-100-LT and 0.05 for ImageNet-LT. In one communication round 20 clients are selected and each client trains 2 local epochs. The number of communication rounds is 2000 for most methods on CIFAR-10/100-LT and 1000 on ImageNet-LT. More details are reported in Appendix.

4.1 RESULTS

Comparison with Previous Methods. To evaluate the effectiveness of our group-based multi-expert model, we conduct experiments comparing it with previous methods including the baseline algorithm FedAvg McMahan et al. (2017), the method for dealing with data heterogeneity FedProx Li et al. (2020b), the methods for dealing with class imbalance in FL Ratio LossWang et al. (2021) and CLIMBShen et al. (2021) , and centralized long-tailed learning methods Focal Loss Lin et al. (2017), CRT-IB, CRT-CB Kang et al. (2019), LDAM Cao et al. (2019), BSM Ren et al. (2020), LADE Hong et al. (2021), RIDE Wang et al. (2020), and Paco Cui et al. (2021). The test accuracy on three typical long-tailed datasets are summarized in Table 1. Our approach significantly outperforms the previous works on each dataset. Notice that the re-balance methods that need prior information utilize our GPI. They have better results than other methods, especially on CIFAR-100-LT and ImageNet-LT, which have a large number of classes. The subsequent experiments show that the GPI makes considerable contribution on their good performance.

Effectiveness of GPI. To further verify the effectiveness of GPI, we use the local label distribution (local re-balance), global label distribution (global re-balance), and GPI as the prior information for

Table 2: Comparison of local re-balance, global re-balance and GPI with different re-balance strategies on CIFAR10-LT, CIFAR100-LT and ImageNet-LT.

Methods		CIFAR-10-LT		CIFAR-100-LT		ImageNet-LT
		100	50	100	50	
Local Re-balance	FedAvg	53.16	61.67	34.48	36.84	33.80
	LDAM	63.55	69.50	32.54	37.01	30.70
	BSM	66.64	74.29	35.55	39.50	34.39
	LADE	66.05	74.54	33.28	39.89	36.32
	RIDE	59.95	69.05	34.94	39.09	40.00
	PaCo	65.11	71.35	35.92	40.03	38.26
Global Re-balance	LDAM	67.69	72.41	36.61	40.01	38.66
	BSM	65.81	74.15	38.56	42.45	39.63
	LADE	66.03	74.06	39.82	42.95	39.37
	RIDE	60.37	68.39	34.59	39.35	40.81
	PaCo	69.60	72.36	37.63	41.26	39.74
	Ours	69.76	75.12	**41.77**	**45.32**	45.57
GPI	LDAM	66.73	72.58	35.36	39.17	34.88
	BSM	67.44	74.36	37.19	41.91	37.64
	LADE	66.89	74.79	38.61	41.08	38.52
	RIDE	58.64	68.63	35.88	39.39	40.30
	PaCo	68.84	75.35	37.76	43.24	39.04
	Ours	**71.07**	**76.85**	40.42	45.16	**45.75**

Table 3: Top-1 accuracy of Many/Medium/Few classes on Cifar-100-LT with IR=100.

Methods	Local Re-balance			Global Re-balance			GPI		
	Many	Medium	Few	Many	Medium	Few	Many	Medium	Few
FedAvg	62.03	32.26	4.93	-	-	-	-	-	-
LDAM	56.23	29.91	7.97	52.03	37.63	17.43	56.14	33.80	12.93
BSM	57.63	36.34	8.87	51.54	41.49	20.00	52.80	39.26	19.07
LADE	54.00	33.74	8.57	52.57	43.34	20.83	55.34	40.46	12.20
RIDE	64.97	31.11	4.37	**63.29**	32.34	3.73	**63.66**	34.91	4.60
PaCo	35.60	46.34	24.13	42.06	41.23	**28.27**	45.94	42.23	22.67
Ours	-	-	-	50.29	**44.71**	28.07	49.97	**44.14**	**24.93**

several typical centralized long-tailed methods. As shown in Table 2, GPI exhibits better performance than local re-balance and comparable results to global re-balance, indicating that our method can handle imbalance issue for FL without data privacy leakage.

Furthermore, we report the accuracy of many/medium/few-shot classes on CIFAR-100-LT dataset with IR=100 in Table 3. Compared with local re-balance using local label distribution, previous cost-sensitive re-balance methods (*i.e.*, LDAM, Balanced SM and LADE) with GPI can perform significant improvements for medium-shot and few-shot classes. Similarly, benefiting from GPI, RIDE and PaCo also further increase the recognition accuracy for medium-shot and many-shot classes. The slight performance improvement of few-shot classes may be attributed to that RIDE and PaCo focus on representation quality over all classes rather than merely minority classes. On the other hand, compared with global re-balance, GPI shows competitive results for all classes, although it exhibits relatively lower accuracy for the medium-shot and few-shot classes. Notice that GPI performs better on many-shot classes sometimes due to the subtraction of the probability term as we mentioned in section 3.2. Compared with the previous SOTA method RIDE, our method obtains higher accuracy since heterogeneous imbalance is alleviated via client grouping.

GPI Visualization. As shown in Figure 3, We visualize the class-wise GPI curves on CIFAR100-LT with IR=10 and IR=100. Compared with the groundtruth label distribution on the same IR, the GPI curves exhibit similar tendency, which can be used for re-balance strategies. Besides, GPI can maintain the performance of majority classes by using smaller GPI values for re-balancing. It should be noticed that GPI is merely obtained from the model updates implicitly combining label frequency and prediction probability as mentioned in Theorem 2.

Ablation Studies on the Ensemble Framework. Table 4 shows the component analysis of the proposed ensemble framework. We compare three different grouping methods: grouping clients randomly (Random Grouping), grouping clients by the cosine distances between local and global label distributions (LD Grouping), and grouping clients by the cosine distances between LPI and GPI (PI Grouping). The results show that Random Grouping contributes little for the performance, while LD Grouping and PI Grouping bring about 3% accuracy improvement. Our grouping method can further improve the final performance of other existing re-balance strategies besides BSM Ren et al. (2020) such as LDAM loss Cao et al. (2019). Moreover, we evaluate the influence of different values of α, the hyper-parameter to control the client selection scope of each expert. As shown in Figure 4, the top-1 accuracy first rises and then falls along with increasing α, which demonstrates that client grouping is useful to learn a balanced global model. The performance can be enhanced via an appropriate value of α.

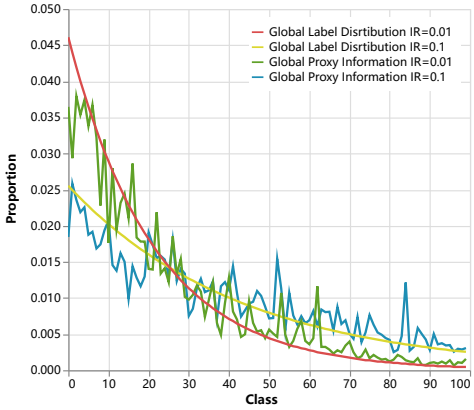


Figure 3: Comparison between the curves of global label distribution and global proxy information on CIFAR100-LT with IR=10 and IR=100.

Methods	Accuracy
LDAM	35.36
BSM	37.19
PI Grouping + LDAM	38.25
Random Grouping + BSM	37.38
LD Grouping + BSM	40.51
PI Grouping + BSM (Ours)	40.42

Table 4: Ablation studies for the grouping method and training loss of the ensemble framework on CIFAR-100-LT with IR=100.

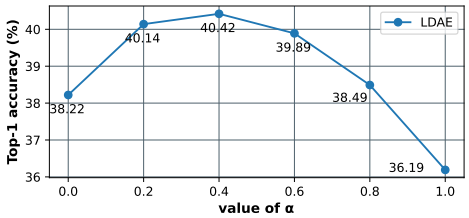


Figure 4: Ablation studies on different α for client selection in the ensemble framework on CIFAR-100-LT with IR=100.

Data Heterogeneity. Existing re-balance strategies under FL setting generally aim to train a balanced local model with compromising performance, which may ignore the heterogeneous property of FL. As shown in Table 1, Table 2 and Table 3, GPI can significantly improve the performance by alleviating the data heterogeneity issue. Moreover, our multi-expert framework could ensemble knowledge learned from different local data sources to further address heterogeneous imbalance.

Privacy Protection. To train a balanced global model, previous re-balance methods mainly rely on global label distribution without privacy protection, while the proposed GPI is merely related to the model updates without extra information. Thus, our method can maintain the data privacy of clients and simultaneously achieve better performance on all the benchmarks.

5 Conclusions

In this work, we focus on the long-tailed distribution problem in FL. Existing re-balance methods for FL suffer from the data heterogeneity issue and assume label distribution is available without considering privacy requirement. To this end, we propose a label-distribution-agnostic ensemble learning (LDAE) framework to re-balance the global objective with preserving privacy. In particular, we design the proxy information to guide the balanced training of multiple experts, which only involves model updates for preserving privacy. Extensive experimental results indicate the effectiveness of our method compared with previous state-of-the-art methods.

REFERENCES

Jiarui Cai, Yizhou Wang, and Jenq-Neng Hwang. Ace: Ally complementary experts for solving long-tailed recognition in one-shot. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 112–121, 2021.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Jiequan Cui, Zhisheng Zhong, Shu Liu, Bei Yu, and Jiaya Jia. Parametric contrastive learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 715–724, 2021.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9268–9277, 2019.

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 Transactions on Parallel and Distributed Systems*, 32(1):59–71, 2020.

Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Youngkyu Hong, Seungju Han, Kwanghee Choi, Seokjun Seo, Beomsu Kim, and Buru Chang. Disentangling label distribution for long-tailed visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6626–6636, 2021.

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.

Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2019.

Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank Reddi, Sebastian Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning*, pp. 5132–5143. PMLR, 2020.

Qinbin Li, Zeyi Wen, Zhaomin Wu, Sixu Hu, Naibo Wang, Yuan Li, Xu Liu, and Bingsheng He. A survey on federated learning systems: vision, hype and reality for data privacy and protection. *IEEE Transactions on Knowledge and Data Engineering*, 2021.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020a.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2:429–450, 2020b.

Wenqi Li, Fausto Milletarì, Daguang Xu, Nicola Rieke, Jonny Hancox, Wentao Zhu, Maximilian Baust, Yan Cheng, Sébastien Ourselin, M Jorge Cardoso, et al. Privacy-preserving federated brain tumour segmentation. In *International workshop on machine learning in medical imaging*, pp. 133–141. Springer, 2019.

Xiaoxiao Li, Meirui JIANG, Xiaofei Zhang, Michael Kamp, and Qi Dou. Fedbn: Federated learning on non-iid features via local batch normalization. In *International Conference on Learning Representations*, 2020c.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.

Quande Liu, Cheng Chen, Jing Qin, Qi Dou, and Pheng-Ann Heng. Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1013–1023, 2021.

Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2537–2546, 2019.

Mohammad Malekzadeh, Burak Hasircioglu, Nitish Mital, Kunal Katarya, Mehmet Emre Ozfatura, and Deniz Gündüz. Dopamine: Differentially private federated learning on medical data. *arXiv preprint arXiv:2101.11693*, 2021.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

Shiva Raj Pokhrel and Jinho Choi. Federated learning with blockchain for autonomous vehicles: Analysis and design challenges. *IEEE Transactions on Communications*, 68(8):4734–4746, 2020.

Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, et al. Balanced meta-softmax for long-tailed visual recognition. *Advances in Neural Information Processing Systems*, 33:4175–4186, 2020.

Sumudu Samarakoon, Mehdi Bennis, Walid Saad, and Mérouane Debbah. Distributed federated learning for ultra-reliable low-latency vehicular communications. *IEEE Transactions on Communications*, 68(2):1146–1159, 2019.

Felix Sattler, Simon Wiedemann, Klaus-Robert Müller, and Wojciech Samek. Robust and communication-efficient federated learning from non-i.i.d. data. *IEEE Transactions on Neural Networks and Learning Systems*, 31(9):3400–3413, 2020. doi: 10.1109/TNNLS.2019.2944481.

Xinyi Shang, Yang Lu, Gang Huang, and Hanzi Wang. Federated learning on heterogeneous and long-tailed data via classifier re-training with federated features. *arXiv preprint arXiv:2204.13399*, 2022.

Li Shen, Zhouchen Lin, and Qingming Huang. Relay backpropagation for effective learning of deep convolutional neural networks. In *European conference on computer vision*, pp. 467–482. Springer, 2016.

Zebang Shen, Juan Cervino, Hamed Hassani, and Alejandro Ribeiro. An agnostic approach to federated learning with class imbalance. In *International Conference on Learning Representations*, 2021.

Pramote Sittijuk and Kriengsuk Tamee. Performance measurement of federated learning on imbalanced data. In *2021 18th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pp. 1–6. IEEE, 2021.

Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S Talwalkar. Federated multi-task learning. *Advances in neural information processing systems*, 30, 2017.

Yanmin Sun, Mohamed S Kamel, Andrew KC Wong, and Yang Wang. Cost-sensitive boosting for classification of imbalanced data. *Pattern recognition*, 40(12):3358–3378, 2007.

Canh T Dinh, Nguyen Tran, and Josh Nguyen. Personalized federated learning with moreau envelopes. *Advances in Neural Information Processing Systems*, 33:21394–21405, 2020.

Lixu Wang, Shichao Xu, Xiao Wang, and Qi Zhu. Addressing class imbalance in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 10165–10173, 2021.

Xudong Wang, Long Lian, Zhongqi Miao, Ziwei Liu, and Stella X Yu. Long-tailed recognition by routing diverse distribution-aware experts. *arXiv preprint arXiv:2010.01809*, 2020.

Liuyu Xiang, Guiguang Ding, and Jungong Han. Learning from multiple experts: Self-paced knowledge distillation for long-tailed classification. In *European Conference on Computer Vision*, pp. 247–263. Springer, 2020.

Chenguang Xiao and Shuo Wang. An experimental study of class imbalance in federated learning. In *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1–7. IEEE, 2021.

Yifan Zhang, Bryan Hooi, Lanqing Hong, and Jiashi Feng. Test-agnostic long-tailed recognition by test-time aggregating diverse experts with self-supervision. *arXiv preprint arXiv:2107.09249*, 2021a.

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

Bo Zhao, Konda Reddy Mopuri, and Hakan Bilen. idlg: Improved deep leakage from gradients. *arXiv preprint arXiv:2001.02610*, 2020.

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.

Boyan Zhou, Quan Cui, Xiu-Shen Wei, and Zhao-Min 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*, pp. 9719–9728, 2020.

Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. *Advances in neural information processing systems*, 32, 2019.