

Towards Federated Long-Tailed Learning

Zihan Chen^{1,2*}, Songshang Liu^{1*}, Hualiang Wang¹, Howard H. Yang¹,
Tony Q.S. Quek² and Zuozhu Liu^{1†}

¹ZJU-UIUC Institute, Zhejiang University, China

²Singapore University of Technology and Design, Singapore

zihan_chen@mymail.sutd.edu.sg, songshang.17@intl.zju.edu.cn, hualiang_wang@zju.edu.cn,
haoyang@intl.zju.edu.cn, tonyquek@sutd.edu.sg, zuozhuliu@intl.zju.edu.cn.

Abstract

Data privacy and class imbalance are the norm rather than the exception in many machine learning tasks. Recent attempts have been launched to, on one side, address the problem of learning from pervasive private data, and on the other side, learn from long-tailed data. However, both assumptions might hold in practical applications, while an effective method to simultaneously alleviate both issues is yet under development. In this paper, we focus on learning with long-tailed (LT) data distributions under the context of the popular privacy-preserved federated learning (FL) framework. We characterize three scenarios with different local or global long-tailed data distributions in the FL framework, and highlight the corresponding challenges. The preliminary results under different scenarios reveal that substantial future work are of high necessity to better resolve the characterized federated long-tailed learning tasks.

1 Introduction

Federated learning (FL) has garnered increasing attentions from both academia and industries, as it provides an approach for multiple clients to collaboratively train a machine learning model without exposing their private data [McMahan *et al.*, 2017; Bonawitz *et al.*, 2019]. This privacy-preserving feature has prevailed FL in a broad range of applications such as the healthcare, finance, and recommendation systems [Andreux *et al.*, 2020; Yang *et al.*, 2020]. The data stem from different sources often exhibits a high level of heterogeneity, e.g., non-IID distribution and/or imbalance in the size, which impedes the FL performance [Li *et al.*, 2020; Wang *et al.*, 2021]. Although several methods have been proposed to circumvent this issue by tackling the drift and inconsistency between the server and clients [Wang *et al.*, 2020a; Karimireddy *et al.*, 2020], the impacts from long-tailed data distribution, which is an extreme case of data heterogeneity and widely exists in the real world data (e.g., healthcare and user behaviors data [Kang *et al.*, 2019; Shang *et al.*, 2022]), has yet been understood.

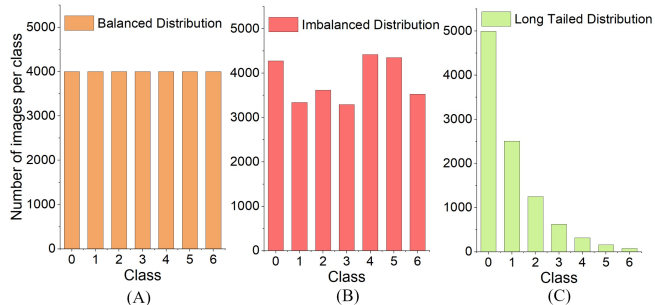


Figure 1: A comparison between the balanced, imbalanced, and long-tailed data distributions over a dataset with 7 classes. (A) is the balanced data distribution. (B) is the imbalanced data distribution. (C) is the long-tailed data distribution.

Unlike data heterogeneity in the general sense, long-tailed distribution has a severely skewed shape in the distribution curve. To better illustrate this phenomenon, we provide a pictorial example in Figure 1. We differentiate the term long-tailed distribution in this figure as well as the rest of this paper to emphasize its unique role. Using Figure 1, we can easily conclude that that in the presence of long-tailed data, training an unbiased classification model is generally challenging since the most of training data is concentrated in a few classes (i.e., the head classes) while the other classes (i.e., the tail classes) have very few samples. And it has been shown in [Kang *et al.*, 2019] that conventional deep learning models admit a significant performance degradation on real-world data that has a long-tailed distribution. In response, several schemes have been proposed to address such an extreme class imbalance issue. These methods are commonly known as the *long-tailed learning*, established via the particular means of re-balancing [Zhang *et al.*, 2021], re-weighting [Lin *et al.*, 2017], and transfer learning techniques [Yin *et al.*, 2019]. Recently, decoupled representation and classification learning scheme [Kang *et al.*, 2019] is investigated to effectively complement the conventional approaches (e.g., class-balanced sampling [Wang *et al.*, 2020b] and distribution-aware loss [Lin *et al.*, 2017]).

However, these existing solutions are primarily dedicated to the centralized learning (CL) and cannot be directly ex-

*Equal contributions. † Corresponding author.

tended to the FL settings. Specifically, due to the distributed nature of the local data, it is much more difficult to train an unbiased model with the existence of long-tailed data in FL systems. Additionally, the limited local dataset sizes of the local clients as well as the inherent data heterogeneity in FL also constrain the applicability of the approaches developed in the scenarios of CL [Yoon *et al.*, 2020].

We refer to the FL task with long-tailed data as the *federated long-tailed learning*. Note that long-tailed data distribution may exist in both the local and global level, leading to different challenges during the training procedure. Particularly, the long-tailed data distribution presents an obvious characteristic on the head and tail over different classes (See Figure 1 (C)). In FL systems, different clients could have different long-tailed properties and the overall (global) data distribution would also be balanced or imbalanced in different networks. The distribution of the real-world datasets is closely related to the user habits and geo-locations, such as the image recognition datasets of the natural specifics (e.g., iNaturalist [Van Horn *et al.*, 2018]) and the landmarks (e.g., Google Landmarks [Weyand *et al.*, 2020]). Such datasets would have a strongly geographical-dominated long-tailed distribution, and more importantly, images from different clients (in different locations) would present different distributional statistics. It would be more challenging to train models with good generalization on different local long-tailed data distributions than the single-distribution case.

Motivated by the aforementioned issues and the intrinsic properties of federated long-tail learning, the present paper gives a comprehensive analysis to the effect of long-tailed data on both the local and global level of FL, as well as the consequent challenges. In addition, numerical results in different settings are also provided to demonstrate the influence of long-tailed data distribution. Based on this, several future trends and open research opportunities are also discussed.

2 Problem Formulation of Federated Long-Tailed Learning

In this section, we will systematically characterize the Federated Long-Tailed (F-LT) learning problem, with the main difference lies at the distributions of the local data in each FL client and the aggregated global data distributions. The challenges under each setting are also discussed in detail.

2.1 Local and global data distribution

Consider an FL system with N clients and an M -class visual recognition dataset for classification problems, where \mathcal{D}_k represents the local dataset for client k . Let n_k denote the size of the local dataset for client k (i.e., $|\mathcal{D}_k|$), and $n_k^{(i)}$ denote the number of data samples of class i in \mathcal{D}_k , i.e., $n_k = \sum_{i=1}^M n_k^{(i)}$.

For a given client k , we shall define the *local data distribution* as

$$\mathbf{p}_k = \left[\frac{n_k^{(1)}}{n_k}, \dots, \frac{n_k^{(j)}}{n_k}, \dots, \frac{n_k^{(M)}}{n_k} \right], \quad (1)$$

where $\frac{n_k^{(j)}}{n_k}$ denotes the ratio of the j -th class over the corresponding local dataset size of client k .

Note that in a typical FL system, the global server does not hold any data. To better capture the overall data distribution from the system level, we define the *global data distribution* as the distribution of the aggregated dataset from all clients in the system, which is denoted by

$$\mathbf{p}_G = \left[\frac{\sum_{k=1}^N n_k^{(1)}}{|\mathcal{D}|}, \frac{\sum_{k=1}^N n_k^{(2)}}{|\mathcal{D}|}, \dots, \frac{\sum_{k=1}^N n_k^{(M)}}{|\mathcal{D}|} \right], \quad (2)$$

where $|\mathcal{D}| = \sum_{k=1}^N n_k$ is the total number of samples in the FL system.

Based on these two length- M vectors \mathbf{p}_k and \mathbf{p}_G , we can illustrate and analyze the distributional statistics of the long-tailed data from both the local and global perspectives. Specifically, the metric *imbalance factor* (IF) [Zhou *et al.*, 2020; Kang *et al.*, 2019] could be used to measure the degree of long-tailed data distribution. Given the local data distribution vector, the local imbalance factor for client k is calculated by

$$\text{IF}_L^{(k)} = \frac{\max_j \{n_k^{(j)}\}}{\min_s \{n_k^{(s)}\}}. \quad (3)$$

Similarly, the global imbalance factor shall be denoted as

$$\text{IF}_G = \frac{\max_j \{\sum_{i=1}^N n_i^{(j)}\}}{\min_s \{\sum_{i=1}^N n_i^{(s)}\}}. \quad (4)$$

2.2 Local and global long-tailed data distribution

Note that either $\text{IF}_L^{(k)}$ or IF_G would be a large number in real-world datasets, which indicates that the long-tailed data distribution may exist in either the local side or global side. For example, the local medical image datasets in hospitals in a big city might follow long-tailed local distributions, while the aggregated city-level global dataset might be long-tailed or non long-tailed. Therefore, considering the relations and differences between the local and global data distributions, we would categorize the federated long-tailed learning tasks into the following three types:

- **Type 1: Both the local and global data distribution follow the same long-tailed distribution.** In a homogeneous network, local data from all the clients follow the same distribution. In such a case, if the local data distribution has the long-tail characteristic, then the global data distribution would also be an identical long-tailed distribution.
- **Type 2: Global data distribution is long-tailed, while local data distributions are diverse, and not necessarily long-tailed.** Local data of different clients in a heterogeneous network would be typically non-IID, where the pattern of the local data distribution would be rarely identical. Given a global long-tailed data distribution, the local data distributions of different clients could be long-tailed, imbalanced or balanced.
- **Type 3: All or a subset of local clients have long-tailed data distributions, but the global data follows a non long-tailed distribution (e.g., balanced distribution over all classes).** In the case that the global data

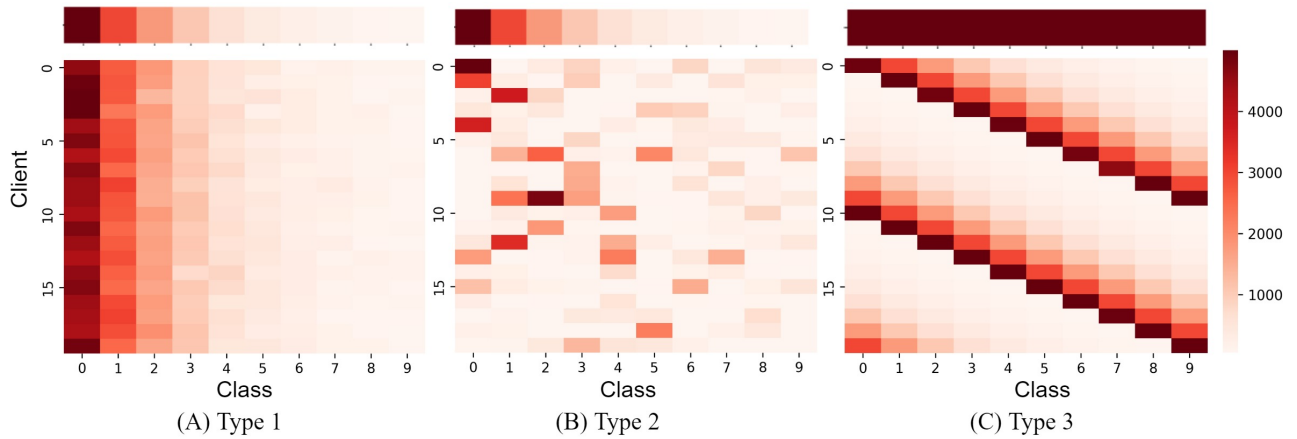


Figure 2: An example of the data distributions for the summarized three types in a 20-clients FL system. The first row is the global data distribution in the corresponding type, with a colorbar in the right indicating the number of data samples in each class. Each sub-colorbox represents the number of data samples of each class across all clients.

Global data distribution	Local data distributions	Objective of learning tasks	Datasets
Long-tailed	Identical long-tailed distributions	Learn a good global model	Long-tailed datasets (e.g., CIFAR-10-LT)
	Long-tailed/ Imbalance/ Balanced distribution	Learn multiple good local models	Long-tailed datasets (e.g., CIFAR-10-LT)
Non long-tailed	Diversified long-tailed distributions	Learn multiple good local models	Balanced datasets (e.g., CIFAR-10)

Table 1: A taxonomy of long-tailed data distribution in FL. The objectives and potential datasets for the corresponding cases in federated long-tail learning are also provided.

distribution is non long-tailed, the pattern of the local long-tailed data distributions of different clients would be diverse (i.e., different clients are supposed to keep different head and tail classes.).

Incorporating the data heterogeneity (i.e., the non-IID and imbalanced dataset size), the overall three cases represent all possible scenarios of long-tailed data in a typical FL system. As illustrated in Figure 2, we provide an example of the summarized three types for better visualization of the local and global distributions in federated long-tailed learning.

2.3 Objective of learning tasks and potential approaches

With the existence of long-tailed data distributions in FL systems, different cases would bring different challenges to the distributed learning process. We will discuss the characterized three types one by one respectively.

In the first type of long-tailed data distribution, local and global data distributions share the same statistical characteristics. A single well-trained global model has the potential to be well generalized over the local data from different clients in FL systems. As the long-tailed distributions of all clients are the same, one classifier trained for long-tailed data could be applicable for all clients. Nevertheless, potential issues may arise due to the limited local dataset sizes.

In the remaining two types, a single distribution could not cover all possible distributions of the clients in the FL sys-

tem. Conventional approaches for long-tail learning for a single long-tailed distribution may fail to tackle such diversity issues. We shall consider different learning objectives for different cases of local and global data distributions. Specifically, different local clients could have vastly diverse distributions (e.g., long-tailed and non long-tailed), and the global and local data distributions would be different. Thus, it is necessary to train multiple models to address such discrepancies of data distributions.

Recall that, in the context of the personalized federated learning (PFL) [Tan *et al.*, 2022], personalized models for each client are trained, as one global model cannot be well generalized to diverse local clients. It would be natural to regard PFL as a key ingredient to tackle such diverse data distribution issues in these two scenarios. For example, a popular solution of PFL is to decouple the local model into base layers and personalization layers [Arivazhagan *et al.*, 2019]. Recent works in the centralized long-tail learning demonstrate that decoupling the representation learning and classifier learning with a re-adjustment on classifier could effectively improve the performance [Kang *et al.*, 2019; Zhou *et al.*, 2020]. Such similar decoupling approaches on model parameters would intuitively make PFL approaches to complement the federated long-tail learning.

From a more general explanation, the key idea of the PFL is to find a good trade-off to balance the global shared knowledge and the local task-specific knowledge for personalized

local training. Such a learning procedure could be applied to learn unbiased long-tail classifiers with a good generalizable representation. Moreover, multi-task learning (MTL) [Smith *et al.*, 2017], clustering [Ghosh *et al.*, 2020] and transfer learning approaches [Gao *et al.*, 2019] could also have the potential to be applied to cross-device long-tail learning in FL, which shall be discussed later in detail (See Sec. 4).

3 Benchmarking the Federated Long-Tailed Learning

To the best of our knowledge, the long-tailed learning in the context of FL has been rarely explored. In this section, we will give a summary on the datasets and the corresponding federated partition approaches. Recent works on long-tail learning in both centralized and federated scenarios will then be discussed. At last, we would give a brief comparison on the two typical long-tailed data settings.

3.1 Datasets and partition methods

Datasets In a centralized paradigm for visual recognition tasks, there are mainly two types of dataset benchmarking for long-tailed study. The first type is the long-tailed version of image datasets modified with synthetic operation, such as exponential sampling (CIFAR10/100-LT [Cao *et al.*, 2019]) and Pareto sampling (ImageNet-LT [Liu *et al.*, 2019], Places-LT [Liu *et al.*, 2019]). They are shaped/sampled from the existing balanced dataset and the degree of the long-tail could be controlled with an arbitrary imbalance factor IF_G . Second type is the real-world large scale datasets with a highly imbalanced label distribution, like iNaturalist [Van Horn *et al.*, 2018] and Google Landmarks [Weyand *et al.*, 2020]. More long-tailed datasets are used in some specific tasks, such as object detection Lvis [Gupta *et al.*, 2019], multi-label classification VOC-MLT [Wu *et al.*, 2020] and COCO-MLT [Wu *et al.*, 2020].

Partition methods for long-tailed FL To create different federated (distributed) datasets according to the different patterns of local and global data distribution, different datasets and sampling methods are required. Data distributions in Type 1 could be realized by IID sampling on long-tailed datasets. Similarly, Type 2 could be achieved by Dirichlet-distribution [Hsu *et al.*, 2019] based generation method on the long-tailed datasets. Specifically, the degree of the long-tail and the identicalness of local data distributions could be controlled by the global imbalance factor IF_G and the concentration parameter α respectively. And Type 3 could be realized via the different long-tailed sampling (different head and tail pattern) on the balanced datasets.

3.2 Approaches

Centralized long-tail learning In the centralized scenario, long-tailed learning seeks to address the class imbalance in training data. The most direct way is to rebalance the samples of different classes during the model training, such as ROS and RUS [Zhang *et al.*, 2021], Simple calibration [Wang *et al.*, 2020b] and dynamic curriculum learning [Wang *et al.*, 2019]. The balancing ideology could also be implemented in

re-weighting and remargining the loss function, such as Focal Loss [Lin *et al.*, 2017], LDAM Loss [Cao *et al.*, 2019]. These class rebalancing methods could improve the tail performance at the expense of head performance.

To address the limitation of information shortage, some studies focus on improving the tail performance by introducing additional information, such as transfer learning, meta learning, and network architecture improvement. In transfer learning, there have been methods FTL [Yin *et al.*, 2019] and LEAP [Liu *et al.*, 2020] transferring the knowledge from head classes to boost the performance in tail classes. In [Shu *et al.*, 2019], meta-learning is empirically proved to be capable of adaptively learning an explicit weighting function directly from data, which guarantees robust deep learning in front of training data bias. Recently, some studies design and improve network architecture specific to long-tailed data. For example, different types of classifiers are proposed to address long-tailed problems, such as τ -norm classifier [Kang *et al.*, 2019] and Causal classifier [Tang *et al.*, 2020].

Federated long-tail learning Yet, the only one related work on federated long-tail learning [Shang *et al.*, 2022] utilized classifier re-training to re-adjust decision boundaries, where the discussion is limited within the global long-tailed distribution with local heterogeneity. Methods for other types of local and global data distribution remain to be further explored.

Nevertheless, in the presence of long-tailed data, the discrepancies among local and global data distributions of different clients in the FL system, could be possibly addressed by the techniques in the federated optimization algorithm, such as dynamic regularization [Acar *et al.*, 2021], diverse client scheduling [Cho *et al.*, 2022] and adaptive aggregation. In addition, as we discussed previously in Sec. 2.3, PFL could be applied in federated long-tailed learning to find a balance between the representation and the classification learning. We shall give a detailed discussion on such explorations to boost the performance of federated long-tailed learning in Sec. 4.

Based on the above discussion about the data distribution, datasets and learning objectives, we summarize them into Table 1. Note that, the case, where both the local and global data distributions are non-long-tailed, is not listed in this table, as this case is not within the scope of this paper.

3.3 Performance comparison

To better illustrate the impacts of the long-tail data distribution, we shall provide some numerical results with different types of long-tailed data distribution in Tables 2 and 3. For all the experiments, we consider a FL with 40 clients. And the non-IID data partition is implemented by Dirichlet distribution. Apart from the baseline FedAvg [McMahan *et al.*, 2017], the other three FL algorithms are FedProx [Li *et al.*, 2020], CReFF [Shang *et al.*, 2022] and FedPer [Arivazhagan *et al.*, 2019], which are representative approaches to tackle data heterogeneity, long-tailed data and personalization in FL respectively.

Note that, the main purpose of this subsection is to analyze the performance of the different FL methods with diverse data settings to provide some possible insights to the design of the federated long-tailed learning algorithm.

Data Setting	Non-LT ($IF_G = 1$)			$IF_G = 10$			$IF_G = 50$			$IF_G = 100$		
	IID	Non-IID		IID	Non-IID		IID	Non-IID		IID	Non-IID	
	-	$\alpha=1$	$\alpha=0.5$	-	$\alpha=1$	$\alpha=0.5$	-	$\alpha=1$	$\alpha=0.5$	-	$\alpha=1$	$\alpha=0.5$
FedAvg	0.9331	0.9257	0.9151	0.8771	0.8661	0.8545	0.7871	0.7756	0.7465	0.7353	0.7253	0.6916
FedProx	0.9372	0.9290	0.9221	0.8741	0.8730	0.8545	0.7751	0.7499	0.7550	0.7105	0.7092	0.6935
CReFF	0.9432	0.9344	0.9262	0.8875	0.8661	0.8493	0.8035	0.7846	0.7264	0.7275	0.7051	0.6921
FedPer	0.9302	0.9214	0.9139	0.8783	0.8694	0.8643	0.7633	0.7503	0.7478	0.7318	0.7016	0.6964

Table 2: Test accuracies of various FL methods on CIFAR-10-LT with different federated data partitions (i.e., Type 2). Results on balanced CIFAR-10 are also provided for reference and comparison.

Local Setting	$IF_L = 10$	$IF_L = 50$	$IF_L = 100$
FedAvg	0.8805	0.8391	0.8361
FedProx	0.8909	0.8551	0.8419
CReFF	0.8974	0.8646	0.8422
FedPer	0.8950	0.8561	0.8421

Table 3: Test accuracies on CIFAR-10 with different local long-tailed distributions (i.e., Type 3).

We choose two typical long-tailed data distributions in the federated setting to evaluate the performance. In Table 2, we give the results on both the IID and non-IID data settings built upon the global long-tailed dataset CIFAR-10-LT with different imbalance factors 10, 50 and 100. For non-IID data partition, we use Dirichlet distribution-based sampling method with different concentration parameter α to control the degree of data heterogeneity. To better demonstrate the impacts of the long-tailed data distribution, we also include a group of experiment results on the (balanced) CIFAR-10 for reference. In Table 3, results on CIFAR-10 are provided, where we consider sample different long-tailed local data distributions (i.e., different head-tail distribution) with the same imbalance factor IF_L . See Figure 2(C) for an overview.

For the results in Tables 2 and 3, best test accuracies of all algorithms present a descending sort pattern from the left to right, as the degree of the long-tail and heterogeneity is increasing. Interestingly, the federated optimization methods FedProx outperforms FedAvg in the non-long-tailed setting, while it tends to underperform with global long-tailed data in some settings. As a specific method to tackle long-tailed data, CReFF can achieve best results among all four algorithms in most of settings, but it has lower accuracy performances than FedProx with more heterogeneous data distribution. With regard to the PFL methods, our preliminary results illustrate that personalization method outperforms in most of the long-tailed data settings, especially in settings of Table 3 (i.e., diverse local long-tailed distributions in Type 2).

The numerical results indicate that, PFL methods have the potential to enhance the performance without any specialized long-tailed learning techniques. More importantly, the preliminary results also demonstrate the feasibility and possibility to re-purpose the federated optimization and PFL methods with centralized long-tailed learning approaches in federated scenarios.

4 Future Trends and Research Opportunities

Based on the above experimental results and discussions of the federated long-tailed learning, we envision the following directions and opportunities towards the robust and communication-efficient federated long-tailed learning algorithms, architectures and analysis.

- **Incorporate PFL ideas for better federated long-tail learning.** As a promising technique, PFL could possibly boost the training performance of federated long-tailed learning with centralized long-tailed learning methods. How to balance the global shared knowledge with local personalized knowledge could be incorporated into the design of the representation learning and classification architectures in federated long-tailed learning. Moreover, it would be promising to explore the incorporation of the model-based and data-based PFL approaches [Tan *et al.*, 2022] with the long-tailed learning.
- **Hierarchical FL architectures.** In the presence of diverse data distributions, we may consider to group clients with similar long-tail distributional statistics into clusters to jointly learn cluster-level personalized models or conduct cluster-level MTL [Sattler *et al.*, 2020]. However, the design of a privacy-preserving clustering method remains to be further investigated.
- **Re-purpose of existing federated optimization methods.** Local long-tailed data distribution could be regarded as an extremely imbalanced case of data heterogeneity. Hence, how to re-purpose the federated optimization algorithm in the presence of the long-tailed data could be further explored. It would be another open question to develop a heterogeneity-agnostic federated optimization framework. Moreover, MTL-based long-tailed learning could also be a potential approach to address the heterogeneous long-tailed distributions in FL.
- **Design better data partition/sampling schemes or more representative datasets.** In addition to the several real-world long-tailed datasets, most of the current work use the long-tailed version of the popular image datasets. Although this method could use the pre-determined imbalance factor IF_G to control the imbalance, it would also discard a large amount of samples when following the widely-used exponential and Pareto sampling methods. Therefore, the degradation of the performance could also be partially attributed to the small dataset size,

especially for scenarios with a larger imbalance factor in federated settings. How to mitigate such negative impacts should be further investigated. Meanwhile, future research could also leverage on real-world scenarios, such as medical images or autonomous cars, to provide more representative and convincing federated long-tailed learning dataset.

5 Concluding Remarks

In this paper, we introduce the federated long-tailed learning task, a general setting motivated by real-world applications but rarely studied in previous research. We characterize three types of F-LT learning settings with diverse local and global long-tailed data distributions. The benchmark results with multiple federated learning architectures suggest that substantial future work is needed for better F-LT. In addition, we highlight the potential techniques and possible trajectories of research towards federated long-tailed learning with real-world data.

References

- [Acar *et al.*, 2021] Durmus Alp Emre Acar, Yue Zhao, Ramon Matas Navarro, Matthew Mattina, Paul N Whatmough, and Venkatesh Saligrama. Federated learning based on dynamic regularization. *arXiv preprint arXiv:2111.04263*, 2021.
- [Andreux *et al.*, 2020] Mathieu Andreux, Jean Ogier du Terail, Constance Beguier, and Eric W Tramel. Siloed federated learning for multi-centric histopathology datasets. In *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*, pages 129–139. Springer, 2020.
- [Arivazhagan *et al.*, 2019] Manoj Ghuhan Arivazhagan, Vinay Aggarwal, Aaditya Kumar Singh, and Sunav Choudhary. Federated learning with personalization layers. *arXiv preprint arXiv:1912.00818*, 2019.
- [Bonawitz *et al.*, 2019] Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloe Kiddon, Jakub Konečný, Stefano Mazzocchi, Brendan McMahan, et al. Towards federated learning at scale: System design. *Proceedings of Machine Learning and Systems*, 1:374–388, 2019.
- [Cao *et al.*, 2019] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Archiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. In *Advances in Neural Information Processing Systems*, 2019.
- [Cho *et al.*, 2022] Yae Jee Cho, Jianyu Wang, and Gauri Joshi. Client selection in federated learning: Convergence analysis and power-of-choice selection strategies. In *Artificial intelligence and statistics*, 2022.
- [Gao *et al.*, 2019] Dashan Gao, Yang Liu, Anbu Huang, Ce Ju, Han Yu, and Qiang Yang. Privacy-preserving heterogeneous federated transfer learning. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 2552–2559. IEEE, 2019.
- [Ghosh *et al.*, 2020] Avishek Ghosh, Jichan Chung, Dong Yin, and Kannan Ramchandran. An efficient framework for clustered federated learning. *Advances in Neural Information Processing Systems*, 33:19586–19597, 2020.
- [Gupta *et al.*, 2019] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019.
- [Hsu *et al.*, 2019] 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.
- [Kang *et al.*, 2019] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. In *International Conference on Learning Representations*, 2019.
- [Karimireddy *et al.*, 2020] 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*, pages 5132–5143. PMLR, 2020.
- [Li *et al.*, 2020] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Amey Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2:429–450, 2020.
- [Lin *et al.*, 2017] 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*, pages 2980–2988, 2017.
- [Liu *et al.*, 2019] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X. Yu. Large-scale long-tailed recognition in an open world. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.
- [Liu *et al.*, 2020] Jialun Liu, Yifan Sun, Chuchu Han, Zhaopeng Dou, and Wenhui Li. Deep representation learning on long-tailed data: A learnable embedding augmentation perspective. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2970–2979, 2020.
- [McMahan *et al.*, 2017] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguerre y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.
- [Sattler *et al.*, 2020] Felix Sattler, Klaus-Robert Müller, and Wojciech Samek. Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints. *IEEE transactions on neural networks and learning systems*, 32(8):3710–3722, 2020.

- [Shang *et al.*, 2022] 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.
- [Shu *et al.*, 2019] Jun Shu, Qi Xie, Lixuan Yi, Qian Zhao, Sanping Zhou, Zongben Xu, and Deyu Meng. Meta-weight-net: Learning an explicit mapping for sample weighting. *Advances in neural information processing systems*, 32, 2019.
- [Smith *et al.*, 2017] Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S Talwalkar. Federated multi-task learning. *Advances in neural information processing systems*, 30, 2017.
- [Tan *et al.*, 2022] Alysa Ziyang Tan, Han Yu, Lizhen Cui, and Qiang Yang. Towards personalized federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.
- [Tang *et al.*, 2020] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect. *Advances in Neural Information Processing Systems*, 33:1513–1524, 2020.
- [Van Horn *et al.*, 2018] Grant Van Horn, Oisín Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018.
- [Wang *et al.*, 2019] Yiru Wang, Weihao Gan, Jie Yang, Wei Wu, and Junjie Yan. Dynamic curriculum learning for imbalanced data classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5017–5026, 2019.
- [Wang *et al.*, 2020a] Jianyu Wang, Qinghua Liu, Hao Liang, Gauri Joshi, and H Vincent Poor. Tackling the objective inconsistency problem in heterogeneous federated optimization. *Advances in neural information processing systems*, 33:7611–7623, 2020.
- [Wang *et al.*, 2020b] Tao Wang, Yu Li, Bingyi Kang, Junnan Li, Junhao Liew, Sheng Tang, Steven Hoi, and Jiashi Feng. The devil is in classification: A simple framework for long-tail instance segmentation. In *European conference on computer vision*, pages 728–744. Springer, 2020.
- [Wang *et al.*, 2021] Lixu Wang, Shichao Xu, Xiao Wang, and Qi Zhu. Addressing class imbalance in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 10165–10173, 2021.
- [Weyand *et al.*, 2020] Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2-a large-scale benchmark for instance-level recognition and retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2575–2584, 2020.
- [Wu *et al.*, 2020] Tong Wu, Qingqiu Huang, Ziwei Liu, Yu Wang, and Dahua Lin. Distribution-balanced loss for multi-label classification in long-tailed datasets. In *European Conference on Computer Vision*, pages 162–178. Springer, 2020.
- [Yang *et al.*, 2020] Liu Yang, Ben Tan, Vincent W Zheng, Kai Chen, and Qiang Yang. Federated recommendation systems. In *Federated Learning*, pages 225–239. Springer, 2020.
- [Yin *et al.*, 2019] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Feature transfer learning for face recognition with under-represented data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5704–5713, 2019.
- [Yoon *et al.*, 2020] Tehrim Yoon, Sumin Shin, Sung Ju Hwang, and Eunho Yang. Fedmix: Approximation of mixup under mean augmented federated learning. In *International Conference on Learning Representations*, 2020.
- [Zhang *et al.*, 2021] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *arXiv preprint arXiv:2110.04596*, 2021.
- [Zhou *et al.*, 2020] 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*, pages 9719–9728, 2020.