Federated Partial Label Learning with Local-Adaptive Augmentation and Regularization

Yan Yan[1], **Yuhong Guo**[1, 2]

[1]School of Computer Science, Carleton University, Ottawa, Canada
[2]Canada CIFAR AI Chair, Amii
yanyan@cunet.carleton.ca, yuhong.guo@carleton.ca

Abstract

Partial label learning (PLL) expands the applicability of supervised machine learning models by enabling effective learning from weakly annotated overcomplete labels. Existing PLL methods however focus on the standard centralized learning scenarios. In this paper, we expand PLL into the distributed computation setting by formalizing a new learning scenario named as federated partial label learning (FedPLL), where the training data with partial labels are distributed across multiple local clients with privacy constraints. To address this challenging problem, we propose a novel Federated PLL method with Local-Adaptive Augmentation and Regularization (FedPLL-LAAR). In addition to alleviating the partial label noise with moving-average label disambiguation, the proposed method performs MixUp-based local-adaptive data augmentation to mitigate the challenge posed by insufficient and imprecisely annotated local data, and dynamically incorporates the guidance of global model to minimize client drift through adaptive gradient alignment regularization between the global and local models. Extensive experiments conducted on multiple datasets under the FedPLL setting demonstrate the effectiveness of the proposed FedPLL-LAAR method for federated partial label learning.

Introduction

Partial label learning (PLL) addresses a weakly supervised learning problem, where each training instance is associated with a noisy set of overcomplete candidate labels, only one of which is the true label (Yu and Zhang 2016; Lyu et al. 2019; Feng et al. 2020; Wang et al. 2022). With the capacity of handling noisy overcomplete labels, PLL expands the applicability of standard supervised learning by substantially reducing its dependence on high quality precise data annotations. Current study on PLL however entirely focuses on a centralized learning scenario where all the training data samples are collected to the training center.

Recently, with the advent of edge computing over distributed data sources across different clients such as personal devices (Konečnỳ et al. 2016), financial institutions (Long et al. 2020), and hospitals (Guo et al. 2021), data privacy and security concerns raise significant demands for decentralized learning methodologies (Kairouz et al. 2021). Federated learning (FL), as a decentralized paradigm of training a global model by enabling collaborations between the individual clients without compromising their data privacy, has hence received increasing attention from the research community (McMahan et al. 2017; Konečnỳ et al. 2016). Nevertheless existing FL methods assume the availability of precise data labels (Li et al. 2020b) that often induces high data annotation cost, and lack the capacity of exploiting weak annotations such as overcomplete partial labels that can often be produced from inexpensive crowdsourcing or high-level text summaries (Feng and An 2018, 2019).

In this paper, we aim to bridge the gap between the existing separate research studies on PLL and FL by formalizing a new learning problem named as federated partial label learning (FedPLL), which extends partial label learning into a federated learning setting where data are distributed across local clients with privacy constraints for data sharing, and the local data is annotated with imprecise partial labels. FedPLL can be degenerated into the standard FL problem when the local data can be precisely annotated. As high quality data annotations often require human expertise in many domains, FedPLL provides a more practical distributed learning scenario by enabling federated learning with imprecise local data annotations. Meanwhile, FedPLL inherits challenges from both PLL and FL, while the entangled challenges can be even more difficult to handle: as data privacy constraints prevent data sharing, the model training process can be more prone to noisy labels with limited data accessibility. Such integrated challenges prevent the direct deployment of the off-the-shelf PLL or FL methods for effectively addressing the FedPLL problem.

To address this new FedPLL problem, we propose a novel federated PLL method with local-adaptive augmentation and regularization mechanisms, named as FedPLL-LAAR, to induce good prediction models across multiple clients with limited communication cost without data sharing. The approach simultaneously handles partial labels and local data insufficiency under a federated learning framework. Specifically, FedPLL-LAAR dynamically disambiguates the noisy candidate labels by consulting the prediction outputs of each local model in a moving-average manner, while the disambiguated labels are in turn used as prediction targets for local model training with MixUp-based local-adaptive data augmentation and local-global gradient alignment regularization. The MixUp-based local-adaptive data augmentation

is designed to alleviate the difficulty posed by insufficient local data and improve the model's robustness against noisy labels. The local-global gradient alignment regularization is proposed to guide the local model training with global information and consequently mitigate the local client drift problem. The main contribution of the paper can be summarized as follows: (1) A new distributed learning scenario FedPLL is proposed, which expands federated learning to exploit data with imprecisely annotated partial labels, and hence substantially reduces the demands for expensive high quality annotations. (2) We pioneer a novel FedPLL-LAAR approach to address the new FedPLL problem, which can simultaneously handle partial labels and local data insufficiency without compromising data privacy. (3) Comprehensive experiments are conducted and the results validate the efficacy of the proposed FedPLL-LAAR approach in tackling the challenges of federated learning with partial labels.

Related Work

Partial Label Learning The main difficulty of PLL lies in how to disambiguate the candidate label sets to identify the true labels. Existing disambiguation methods for PLL can be roughly divided into two categories: average-based methods and identification-based methods. For the average-based methods (Hüllermeier and Beringer 2006; Cour, Sapp, and Taskar 2011; Zhang and Yu 2015), all candidate labels for each instance are treated equally as true labels, and the final prediction is made by averaging the classifier outputs within the candidate label set. This type of methods however often yield suboptimal performance by ignoring the difference among the candidate labels. Given the limitations of the average-based approaches, the identification-based strategy emerges as an alternative solution for addressing the candidate labels with discrimination. Following this strategy, some methods consider the true label as a latent variable which can be refined iteratively within the potential labels (Jin and Ghahramani 2002; Nguyen and Caruana 2008; Liu and Dietterich 2012; Yu and Zhang 2016). For example, Jin and Ghahramani (2002) and Nguyen and Caruana (2008); Yu and Zhang (2016) optimize the objective function based on the maximum likelihood criterion and the maximum margin criterion respectively. In the past decade, identification-based methods (Feng and An 2018, 2019; Lyu et al. 2019, 2020; Wu, Wang, and Zhang 2022) have gained more attention and proven to be more effective than average-based methods. While these works achieve good results, they are limited by factors such as being restricted to linear models and facing challenges in handling large-scale datasets.

Recently, more researchers have started studying PLL in the deep learning paradigm to address the above-mentioned limitations. Yao et al. (2020) exploited deep convolutional neural networks to train a deep PLL model with regularization and temporal-ensemble techniques. Recently, pseudo-labeling methods (Yan and Guo 2020; Feng et al. 2020; Lv et al. 2020; Wen et al. 2021; Xu et al. 2021; Wang et al. 2022; Zhang et al. 2022) have received much attention due to their effectiveness in candidate label disambiguation. In particular, Xu et al. (2021) exploited variational inference to gradually improve the distribution of latent labels by assuming that the noisy labels within the candidate label set are often instance dependent. Wang et al. (2022) proposed to update the pseudo-targets by using contrastively learned class-prototypes, which demonstrates remarkable performance on multiple image classification benchmarks.

Nevertheless, all of these works are centralized PLL methods which assume all the training data are accessible without considering distributed data storage and data privacy. This motivates us to investigate the distributed PLL problem.

Federated learning FL enables multiple clients to work together in a distributed learning paradigm for inducing a good global model without sharing their individual data (McMahan et al. 2017; Yang et al. 2019; Li et al. 2020a). FedAvg is a popular FL algorithm introduced in a pioneering FL study (McMahan et al. 2017), which is the first synchronous algorithm specifically designed for the distributed learning scenario. Thereafter Li et al. (2020b) proposed a FedAvg based FL approach FedProx, which aims to alleviate statistical heterogeneity and enhance stability in FL by incorporating a proximal term in local training, while Acar et al. (2021) adopted a dynamically proximal term based on the selected clients. In addition, a number of classical FL methods (Karimireddy et al. 2020; Hsu, Qi, and Brown 2019; Reddi et al. 2020) have placed their emphasis on addressing client discrepancy problems in non-IID data distributions. Regardless of such great research progress on FL, these FL methods assume the local data on each client are precisely labeled, which induces expensive or even impractical annotation cost in some real-world scenarios where expertise for accurate annotation is sparse.

Recently, there has been increasing attention in federated learning with noisy labels (FedNL) due to its effectiveness for reducing annotation costs. Yang et al. (2022) proposed to swap both class representations of local data and local models, which however may pose a threat to data privacy since the raw images can be reconstructed from the class representation (Zhao et al. 2020). Chen et al. (2020); Tuor et al. (2021) tried to address the FedNL problem by deploying shared clean data on the server. FedCorr (Xu et al. 2022) deploys a multi-stage framework for addressing heterogeneous label noise in FL with adaptive local proximal regularization. RoFL (Yang et al. 2022) updates the pseudo-labels by using additional shared information between local clients and server, which also poses a risk of privacy leakage. Han et al. (2022) proposed to handle FL problem by reducing the divergence in feature representations among different clients arising from heterogeneous local training processes.

Different from these works above, the new FedPLL scenario we proposed handles FL with overcomplete noisy candidate labels. We aim to address the FedPLL problem without any data sharing in order to preserve local data privacy.

Problem Formulation

We formalize the new distributed learning scenario of FedPLL as follows. We consider a federated learning system with training data $D = \{D^m\}_{m=1}^{M}$ distributed across M local clients, where $D^m = \{(\mathbf{x}_i^m, \mathbf{y}_i^m, S_i^m)\}_{i=1}^{n^m}$ denotes the local dataset with n^m training instances on the m-th client,

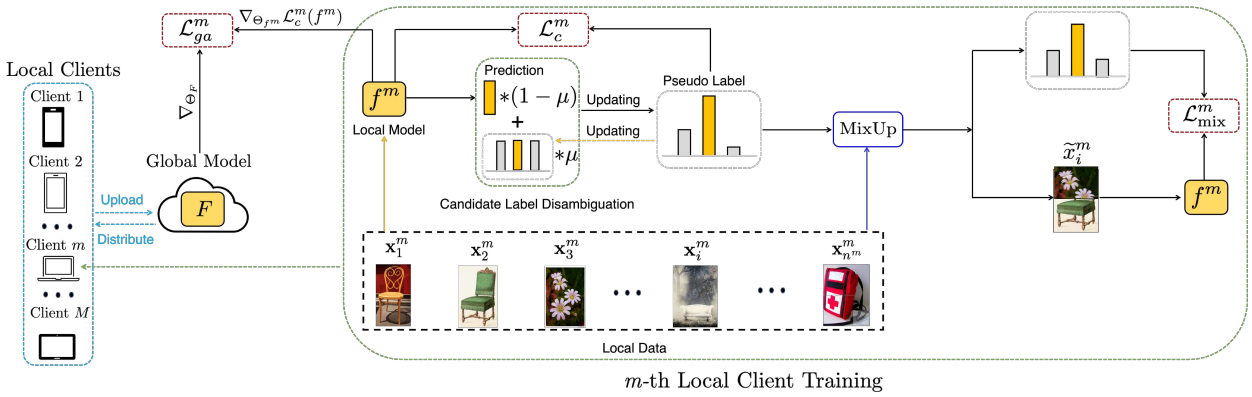

Figure 1: The proposed FedPLL-LAAR model, which has the following three main components: local model learning with disambiguated pseudo-labels ($\mathcal{L}_c^m$); local-global gradient alignment regularization ($\mathcal{L}_{ga}^m$); MixUp-based local adaptive data augmentation ($\mathcal{L}_{\text{mixup}}^m$). All the components cooperate and benefit with each other to address FedPLL.

and $\mathbf{y}_i \in \{0,1\}^L$ denotes the candidate label set indicator vector for instance $\mathbf{x}_i$, which has multiple 1 values corresponding to the true label and additional indistinguishable noise labels in the candidate label set S_i^m. The goal of FedPLL is to induce a good global classification model F by aggregating the M local models $\{f^m\}_{m=1}^M$ trained on their corresponding partially annotated local datasets $\{D^m\}$. The overall FL objective can be written as follows:

$$\min_{\Theta_{f^m}} \mathcal{L} \triangleq \sum_{m=1}^{M} \frac{n^m}{N} \mathcal{L}^m(f^m, D^m) \text{ with } N = \sum_{m=1}^{M} n^m. \quad (1)$$

Here $\mathcal{L}^m(f^m, D^m)$ denotes the local classification loss over the partial label dataset D^m on the m-th client, such as

$$\mathcal{L}^m(f^m, D^m) = \frac{1}{n_m} \ell(\mathbf{y}_i^m, f^m(\mathbf{x}_i^m)), \quad (2)$$

where $\ell(\cdot,\cdot)$ denotes the cross-entropy loss function.

Given the distributed training data, the typical FL framework such as FedAvg works in the following manner by communicating between a global model on the server and local models on the clients in multiple rounds. In each round, parameters of each local model, Θ_{f^m}, are initialized by the global model parameters Θ_F and then updated by minimizing the local classification loss $\mathcal{L}^m(f^m, D^m)$ over the local training data with multiple epochs of gradient descent:

$$\Theta_{f^m} = \Theta_{f^m} - \lambda \nabla_{\Theta_{f^m}} \mathcal{L}^m(f^m, D^m), \quad (3)$$

where λ denotes the local learning rate. Then the global model can be updated by aggregating the local models uploaded to the server:

$$\Theta_F = \sum_{m=1}^{M} \frac{n_m}{N} \Theta_{f^m}. \quad (4)$$

After the global model update, the server distributes its model parameters Θ_F to each local client as the starting point for the next round of local updates.

In this work, we investigate FedPLL within the aforementioned general federated learning framework. As the local training data is annotated with partial labels and often insufficient, the local model can be more prone to overfitting on the sparse and noisy local data, resulting in poor generalizability, especially when the data are non-IID across multiple local clients and form a client drift problem. Moreover, the drifted local models will negatively impact the global model, which further impairs the local models by distributing Θ_F in a negative communication cycle. We address these problems by designing novel local training mechanisms.

Proposed Approach

In this section, we propose a federated partial label learning approach named FedPLL-LAAR to address the entangled challenges of FL and PLL such as local data insufficiency, client drift, and noisy partial labels by designing effective label disambiguation, local-adaptive data augmentation, and local-global regularization mechanisms. The proposed FedPLL-LAAR model is illustrated in Figure 1. It has three pseudo-label based components designed for addressing the FedPLL problem: (1) *local model learning with disambiguated pseudo-labels*, which trains the local model with the disambiguated pseudo-labels that are produced from the candidate labels in a moving average manner; (2) *local-global gradient alignment regularization*, which guides the local model training to mitigate client drift by enforcing gradient alignment between the local and global models; (3) *MixUp-based adaptive local data augmentation*, which performs adaptive local data augmentation through MixUp to alleviate data insufficiency and improve robustness to label noise. All these components collaborate with each other in the aforementioned FedAvg framework to address the FedPLL problem. We elaborate these components and the proposed approach in the following subsections.

Local Model Learning with Disambiguated Pseudo-labels

As the ground-truth labels are concealed in the candidate label sets of the local training data, disambiguating the candidate label sets to identify the true labels is critical for in-

ducing good prediction models. We propose to tackle this label disambiguation challenge by gradually aggregating statistical information from the training data across all clients through the local prediction models.

In particular, on the m-th client we maintain a pseudo-label vector $\mathbf{q}_i^m$ for each training instance $\mathbf{x}_i^m$, and gradually update it towards the true label indicator vector based on the prediction result of the local model f^m in a moving average manner. $\mathbf{q}_i^m$ is initialized by normalizing the corresponding candidate label indicator vector $\mathbf{y}_i^m$, such that $\mathbf{q}_i^m = \frac{1}{|S_i^m|}\mathbf{y}_i^m$, where S_i^m denotes the candidate label set associated with instance $\mathbf{x}_i^m$ and $|S_i^m|$ denotes the number of labels in S_i^m. Then in each epoch of the local model training, we integrate the predictions of f^m to update the pseudo-label vectors for all local instances in D^m in a moving average manner, such that

$$\begin{aligned} \mathbf{p}_{ik}^m &= \mathbb{I}\big[k = \text{argmax}_{k' \in S_i^m} f_{k'}^m(\mathbf{x}_i)\big], \\ \mathbf{q}_i^m &= \mu \mathbf{q}_i^m + (1-\mu)\mathbf{p}_i^m, \end{aligned} \tag{5}$$

where $\mathbb{I}[\cdot]$ is an indicator function that returns value 1 only if the condition within the bracket is true; $f^m(\mathbf{x}_i)$ produces a prediction probability vector over all the classes for instance $\mathbf{x}_i$; $\mu \in (0,1)$ is a momentum hyperparameter that controls the relative amount of information to aggregate from the prediction of the current local model f^m. By uniformly initializing the pseudo-label vectors among the given candidate labels, we start label disambiguation without any prior bias. With the progress of the local model training, the updated pseudo-labels are expected to be more aligned with the true labels—i.e., disambiguated by aggregating statistical information from the trained prediction model.

The disambiguated pseudo-label vectors $\{\mathbf{q}^m\}$ can then in turn be used as targets for further local model training. Specifically, we update the local model parameter Θ_{f^m} by minimizing the following cross-entropy loss on the disambiguated local data:

$$\mathcal{L}_c^m(f^m) = -\frac{1}{n^m}\sum_{i=1}^{n^m} \mathbf{q}_i^{m\top} \log(f^m(\mathbf{x}_i^m)) \tag{6}$$

With mutual pseudo-label disambiguation and local model update, we expect the pseudo-labels can largely converge towards the true labels and consequently push the local model to make correct predictions.

Local-Global Gradient Alignment Regularization

In FedPLL, data is collected and stored locally on different personal devices, such as smartphones and IoT devices. As these devices may have different user behaviors, contexts, and usage patterns, each client often has non-IID (Independent and Identically Distributed) data. In the highly skewed scenario, each client may only have data from a small subset of the classes. Due to such client drift problem, significant divergence exists between the local data distributions and the global data distribution. Simply training each local model on the non-IID local data will largely deviate the local models from the global target, leading to severe local overfitting. To address this challenge, we propose a novel local-global gradient alignment regularization term to reduce client drift by minimizing the gradient dissimilarity between the global and local models. Specifically, we deploy the following gradient alignment loss as a regularizer for the local model f^m given the gradient of the global model F:

$$\begin{aligned} \mathcal{L}_{ga}^m(f^m, F) &= -\text{cosine}\left(\nabla_{\Theta_{f^m}} \mathcal{L}_c^m(f^m), \nabla_{\Theta_F}\right), \\ \nabla_{\Theta_F} &= \sum_{m=1}^{M} \frac{n^m}{N} \nabla_{\Theta_{f^m}} \mathcal{L}_c^m(f^m) \end{aligned} \tag{7}$$

where ∇_{Θ_F} is the gradient of the global model, which is produced by aggregating the gradients of the local models uploaded to the server. $\mathcal{L}_c^m(\cdot)$ is the local classification loss function defined in Eq.(6) over the disambiguated local data on the m-th client using the given prediction model. We expect this local-global gradient alignment regularizer to alleviate the deviation between local model updates and the global objective, preventing overfitting to skewed local data. Moreover, as the global model is updated by aggregating the uploaded local models from all clients, it can be less affected by the local label noise. Thus such a local-global alignment can also enhance the local model training.

Local-Adaptive Data Augmentation

The local training data on each client not only can be non-IID, but also can be very sparse due to the data collection or annotation costs. To address the training data insufficiency on each local client, we propose to conduct MixUp based data augmentation. MixUp is a data augmentation technique developed in (Zhang et al. 2017), which produces synthetic instances by linearly mixing the features and corresponding labels of a given pair of instances, and has demonstrated its robustness on training model against noisy labels (Jiang et al. 2020). Specifically, MixUp produces an augmenting instance $(\widetilde{\mathbf{x}}, \widetilde{\mathbf{y}})$ by linearly interpolating a pair of randomly chosen instances $(\mathbf{x}_i, \mathbf{y}_i)$ and $(\mathbf{x}_j, \mathbf{y}_j)$, such that

$$\widetilde{\mathbf{x}} = \gamma \mathbf{x}_i + (1-\gamma)\mathbf{x}_j, \quad \widetilde{\mathbf{y}} = \gamma \mathbf{y}_i + (1-\gamma)\mathbf{y}_j \tag{8}$$

where $\gamma \sim \text{Beta}(\alpha, \alpha)$, for $\alpha \in (0, \infty)$.

We adopt the MixUp as a data augmentation strategy to alleviate the difficulty of local data sparsity, and further improve the robustness of the local model training against the label noise. Given the local data with disambiguated pseudo-labels, we generate the same amount of augmenting instances by randomly mixing up pairs of instances:

$$\widetilde{\mathbf{x}}_i^m = \gamma \mathbf{x}_i^m + (1-\gamma)\mathbf{x}_{\bar{i}}^m, \quad \widetilde{\mathbf{q}}_i^m = \gamma \mathbf{q}_i^m + (1-\gamma)\mathbf{q}_{\bar{i}}^m \tag{9}$$

where $(\mathbf{x}_{\bar{i}}^m, \mathbf{q}_{\bar{i}}^m) \in \bar{D}^m$ is a dataset after random order shuffling on D^m, and $\mathbf{x}_{\bar{i}}^m$ is the i-th instance from $\bar{D}^m$. For batch-wise training, the operation can be conducted on each training batch instead.

Existing MixUp-based methods treat the mixup instances equally for prediction model training. However, this commonly used strategy can be suboptimal as different instances contribute differently to the prediction model. In light of this, we propose to adaptively weight the augmenting instances for local model training. Specifically, we consider two types of weights for each augmenting instance $\widetilde{\mathbf{x}}_i^m$ based

on the global prediction model. First, we combine the prediction confidence scores on the instance pair $(\mathbf{x}_i^m, \mathbf{x}_{\tilde{i}}^m)$ produced by the global model F using the same linear MixUp interpolation as in Eq.(9) to compute the weight for $\widetilde{\mathbf{x}}_i^m$:

$$w_1(\widetilde{\mathbf{x}}_i^m) = \gamma \cdot \max(F(\mathbf{x}_i^m)) + (1-\gamma) \cdot \max(F(\mathbf{x}_{\tilde{i}}^m)). \quad (10)$$

Then we use the direct prediction confidence score of the global model on $\widetilde{\mathbf{x}}_i^m$ as the second weight:

$$w_2(\widetilde{\mathbf{x}}_i^m) = \max(F(\widetilde{\mathbf{x}}_i^m)). \quad (11)$$

Finally, we combine these two types of weights together in a product form to produce the final weight:

$$w(\widetilde{\mathbf{x}}_i^m) = w_1(\widetilde{\mathbf{x}}_i^m) \cdot w_2(\widetilde{\mathbf{x}}_i^m) \quad (12)$$

The generated augmenting instances can then be incorporated into local model training with the following adaptive MixUp loss—i.e., a weighted cross-entropy loss on the mixup data:

$$\mathcal{L}_{\text{mix}}^m(f^m) = -\frac{1}{n^m} \sum_{i=1}^{n^m} w(\widetilde{\mathbf{x}}_i^m)\, \widetilde{\mathbf{q}}_i^{m\top} \log(f^m(\widetilde{\mathbf{x}}_i^m)) \quad (13)$$

MixUp can generate instances that do not exist in the original local data, expecting to alleviate the non-IID client drift problem. By giving higher weights to instances that are more confidently predicted by the current global model, the loss above can adaptively guide the local model training to incorporate more augmenting information and better align with the global model. Moreover, shifting more attention to confidently predicted instances can also help mitigate label noise.

Overall Training Loss

By integrating the classification loss in Eq.(6), the gradient alignment loss in Eq.(7) and the adaptive MixUp loss in Eq.(13) together, we have the following overall local model training loss:

$$\mathcal{L} = \mathcal{L}_c^m + \eta \mathcal{L}_{ga}^m + \delta \mathcal{L}_{\text{mix}}^m \quad (14)$$

where η and δ are trade-off hyperparameters that control the relative importance of the gradient alignment loss and the adaptive MixUp loss respectively. We minimize this objective for local model training using a mini-batch based stochastic gradient descent (SGD) algorithm.

Experiment

Experiment Setting

Datasets We conducted experiments on three benchmark datasets: (1) CIFAR-10 (Krizhevsky, Hinton et al. 2009) includes 10 categories, and each of them comprises 6,000 images with a resolution of 32 × 32 pixels. (2) CIFAR-100 (Krizhevsky, Hinton et al. 2009) includes 100 categories, each containing 600 images with a resolution of 32 × 32 pixels. (3) SVHN (Netzer et al. 2011) contains a total of 73,257 images for training and 26,032 images for testing, and each of them is a cropped digit of 32 × 32 pixels. In order to utilize the three datasets in the new FedPLL learning scenario, we generate partially annotated non-IID data through the following PLL and FL procedures: (1) For partial labels, we manually corrupt each dataset into a partially labeled version following the partial label generation procedural (Lv et al. 2020). Specifically, any irrelevant label for an image is uniformly selected as a candidate label with a probability ρ, which controls the noise level. When no irrelevant label is chosen for an image, we randomly select an irrelevant label to add to the candidate label set to ensure that the entire training set is thoroughly corrupted. (2) For non-IID data, we followed the procedure of Xu et al. (2022). We use a client-class indicator matrix Φ with size $M \times L$, where L is total number of classes, and each entry Φ_{ij} denotes whether the local data of the i-th client contains the j-th class. The entries of Φ can be sampled from a Bernoulli distribution with a fixed probability σ. For any $j \in [1, L]$, let u_j be the sum of entries in $\Phi[:, j]$, which denotes the number of clients whose local data contain the j-th class. To distribute data across clients, we introduce a vector v_j with length u_j, and sample its values from a symmetric Dirichlet distribution with a parameter β. Then based on the probability vector v_j, we can distribute the instances in the j-th class to the u_j number of local clients recorded in $\Phi[:, j]$.

Comparison Methods We compare the proposed FedPLL-LAAR approach with three state-of-the-art FL methods and one baseline method: (1) FedCorr (Xu et al. 2022), which proposes a multi-stage framework to tackle the heterogeneous noisy label problem by employing adaptive local proximal regularization. (2) RoFL (Yang et al. 2022), which updates the label of uncertain samples by utilizing a global model that works in conjunction with local models through the exchange of class-wise centroids. (3) FedX (Han et al. 2022), which addresses the FL problem with a two-sided knowledge distillation. (4) FedAvg (McMahan et al. 2017) is one of the most widely used baseline FL method, which simply aggregates the parameters of the local models in an element-wise fashion. To ensure a fair comparison, we utilize the same backbone network and classification model for all the methods. Moreover, the comparison methods are configured with the recommended parameters stated in their corresponding works. For each experiment, we report the average test accuracy and the standard deviation based on five independent runs.

Implementation Details In the experiment, we distribute the FedPLL dataset among 10 clients, each of whom possesses their own non-IID local data. The model is trained for 200 communication rounds, with 5, 10, and 10 local epochs in each round on CIFAR-10, SVHN, and CIFAR-100 respectively. We adopt ResNet-18 as the backbone network on CIFAR-10 and SVHN, and adopt ResNet-34 on CIFAR-100. Both the local model f^m on the m-th client and the global model F on the server have the same network structure with a backbone network followed by a liner classifier. The moving average hyperparameter μ for pseudo-label updating and the parameter α used for Beta distribution are set to 0.99 and 0.75 respectively. The probability σ and the parameter β used for producing non-IID data are set to 0.7 and 0.5 respectively. For optimization, we adopt a stochastic gradient descent optimizer with a momentum of 0.5. The learning

Datasets	Setting	Method	$\rho = 0.1$	$\rho = 0.3$	$\rho = 0.5$
CIFAR-10	Centralized	FedPLL-LAAR(Ours)	**94.88±0.15%**	**92.61±0.18%**	**84.99±0.26%**
	Federated	FedPLL-LAAR(Ours)	**93.09±0.20%**	**90.35±0.42%**	**77.66±0.52%**
		FedCorr (Xu et al. 2022)	85.20±0.25%	73.56±0.84%	60.76±0.65%
		RoFL (Yang et al. 2022)	71.11±0.31%	67.99±0.24%	43.83±0.28%
		FedX (Han et al. 2022)	71.97±0.22%	62.84±0.33%	55.80±0.27%
		FedAvg (McMahan et al. 2017)	65.39±0.19%	48.48±0.00%	33.86±0.18%
SVHN	Centralized	FedPLL-LAAR(Ours)	**97.93±0.09%**	**97.51±0.11%**	**96.94±0.06%**
	Federated	FedPLL-LAAR(Ours)	**96.13±0.72%**	**95.81±0.56%**	**93.65±0.36%**
		FedCorr (Xu et al. 2022)	95.29±0.13%	92.92±0.25%	80.16±0.28%
		RoFL (Yang et al. 2022)	91.69±0.14%	88.69±0.46%	65.96±0.34%
		FedX (Han et al. 2022)	90.45±0.05%	85.65±0.09%	79.64±0.13%
		FedAvg (McMahan et al. 2017)	86.91±0.44%	75.92±1.26%	68.14±1.27%

Table 1: Comparisons of accuracy on CIFAR-10 and SVHN with non-IID setting at different noise levels.

Setting	Method	$\rho = 0.01$	$\rho = 0.05$	$\rho = 0.1$
Centralized	FedPLL-LAAR(Ours)	**76.13±0.32%**	**71.39±0.33%**	**64.84±0.23%**
Federated	FedPLL-LAAR(Ours)	**73.27±0.15%**	**68.53±0.19%**	**60.66±0.10%**
	FedCorr (Xu et al. 2022)	59.08±0.21%	48.76±0.29%	35.27±0.23%
	RoFL (Yang et al. 2022)	60.41±0.21%	52.37±0.31%	39.66±0.42%
	FedX (Han et al. 2022)	40.83±0.18%	33.39±0.25%	23.52±0.22%
	FedAvg (McMahan et al. 2017)	34.88±0.23%	26.32±0.26%	20.36±0.12%

Table 2: Comparisons of accuracy on CIFAR-100 with non-IID setting at different noise levels.

rate and mini-batch size are set to 0.03 and 128 respectively in all our comparison experiments. The hyperparameters η and δ used in Eq.(14) are set to 1 by default.

Comparison Results

We compared the proposed FedPLL-LAAR method with three state-of-the-art FL methods and one commonly used FL baseline on three benchmark datasets corrupted at various noise levels in the non-IID FL setting. Table 1 shows the comparison results on CIFAR-10 and SVHN. We can see that FedPLL-LAAR yields the best test accuracies on both datasets across different noise levels. Moreover, the performance gains produced by FedPLL-LAAR are quite remarkable compared with the other comparison methods. For example, FedPLL-LAAR outperforms the best alternative comparison method by 0.84%, 2.89%, and 13.49% when the noise levels are set to 0.1, 0.3, and 0.5 respectively on SVHN. Moreover, FedPLL-LAAR outperforms the best state-of-the-art method by 7.89%, 16.79%, and 16.9% with the noise levels of 0.1, 0.3, and 0.5 respectively on CIFAR-10. It is also worth noting that with the increase of the noise level, the performance gains become larger, which clearly indicates that the proposed FedPLL-LAAR method can effectively handle partial label noise in the federated learning scenario. In addition, we also implemented the proposed approach in centralized learning (CL) to show the performance degradation caused by FL. The accuracy of FedPLL-LAAR used in CL can be regarded as an upper bound for the proposed approach in FedPLL. We can see that the accuracy gaps between the FedPLL-LAAR in CL and FL are relatively small, which demonstrates the efficacy of the proposed approach in addressing FL with challenging non-IID data. In summary, the comparison results reported in Table 1 demonstrate the effectiveness of the proposed FedPLL-LAAR method in addressing FedPLL problems.

Table 2 reports the comparison results on a more complicated and challenging dataset, CIFAR-100, with a larger label space ($L = 100$). It is worth noting that FedPLL-LAAR consistently outperforms all the other comparison methods and improves upon the best comparison method by 12.86%, 16.16%, and 21% with noise levels of 0.01, 0.05, and 0.1 respectively on CIFAR-100. This again demonstrates the efficacy of FedPLL-LAAR in addressing the FedPLL problem.

Impact of Client Numbers To illustrate how would the number of clients affect the performance of FL methods, we plotted the experimental results on the corrupted CIFAR-10 dataset with noise level $\rho = 0.3$ by varying the number of clients in Figure 2. Note that a larger number of clients means that the training data would be distributed on more clients, leading to more severe deviation between the local model updates and the global objective due to the non-IID setting. As shown in Figure 2, we can see that as the number of clients increases, which makes the deviation more severe, the performance degradation of each comparison method becomes more profound. Nevertheless, FedPLL-LAAR consistently outperforms all the other state-of-the-art methods with large performance gains. It is worth noting that FedPLL-LAAR still maintains reasonable prediction performance with a very large number of clients; its classification accuracy exceeds 80% when training data is distributed to 80 clients. This again demonstrates the efficacy of the proposed FedPLL-LAAR in FedPLL.

Communication Cost Communication costs have been a longstanding challenge in FL due to the limitations in existing communication channels. To demonstrate the impact of communication costs on the performance of FL compar-

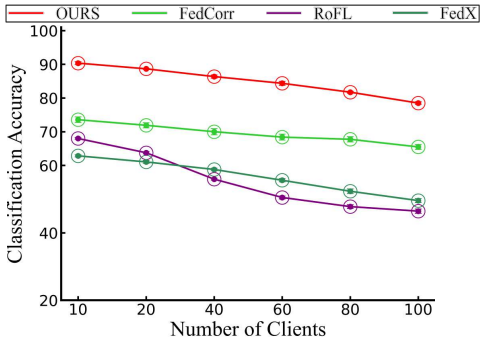

Figure 2: Impact of the client numbers on FedPLL-LAAR and other comparison methods on CIFAR-10.

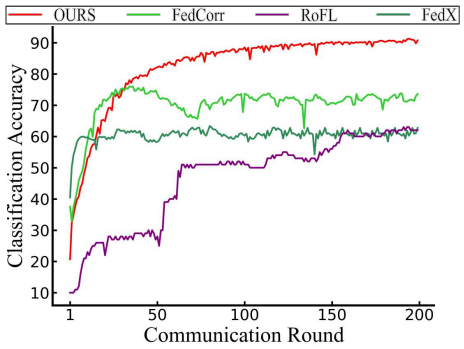

Figure 3: Impact of the communication round on FedPLL-LAAR and other comparison methods on CIFAR-10.

ison methods, we have plotted the results of classification accuracy versus communication rounds on CIFAR-10 in the FedPLL setting with a noise level $\rho = 0.3$ in Figure 3. We can see that FedPLL-LAAR requires fewer communication rounds to converge, while the other methods exhibit oscillations in their prediction results. Moreover, FedPLL-LAAR yields remarkable performance improvements compared to all other methods in the comparison, and these gains persist as the number of communication rounds increases. Although FedPLL-LAAR has a higher communication cost in each round than FedCorr and FedX due to the additional interchange of gradients, the accuracy produced by FedPLL-LAAR with only 50 communication rounds is even higher than the accuracies produced by the other comparison methods across the 200 communication rounds. Hence, overall, FedPLL-LAAR is a more efficient FedPLL method than the others, requiring fewer communication rounds and incurring smaller communication cost to be effective.

Ablation Study

As shown in Eq.(14), the objective of FedPLL-LAAR contains three loss terms: candidate label disambiguation based classification loss, gradient alignment loss, and adaptive MixUp loss. To validate the effectiveness of these components, we perform an ablation study by comparing FedPLL-LAAR with the following ablation variants: (1) CLS-w/o-cld, which drops the candidate label disambiguation; (2) CLS-w/o-am, which drops the adaptive MixUp loss; (3) CLS-w/o-ga, which drops the gradient alignment loss; and (4) CLS-w/o-am-ga, which drops both the adaptive MixUp loss and the gradient alignment loss. Moreover, we further investigate the impact of the adaptive weights for the MixUp loss by considering the following variants: (5) CLS-w/o-w_1, which drops w_1; (6) CLS-w/o-w_2, which drops w_2; and (7) CLS-w/o-$(w_1 + w_2)$, which drops both adaptive weights. The comparison results are reported in Table 3. We can see that compared to the full model, all the ablation variants experience performance degradation to varying degrees, implying that each component contributes to the model's efficacy. Among the seven variants, CLS-w/o-cld leads to the most significant performance degradation. This finding verifies the indispensability of the candidate label disambiguation procedure to FedPLL-LAAR in addressing partial label noise. CLS-w/o-am also produces remarkable performance degradation, which suggests that the adaptive MixUp process can alleviate the challenge posed by sparse local data and enhance the local model's robustness against incorrect pseudo-labels. Among the three adaptive weight variants, CLS-w/o-$(w_1 + w_2)$ yields inferior performance compared to variants CLS-w/o-w_1 and CLS-w/o-w_2, validating the efficacy of the proposed integrated weighting scheme. In addition, the CLS-w/o-ga and CLS-w/o-wm-ga also yield inferior performance, which suggests that the gradient alignment regularization is efficient in alleviating the deviation of local model updates from the global objective. In summary, all the ablation results demonstrate that the proposed FedPLL-LAAR approach is effective in addressing FedPLL by integrating these components.

Ablation variant	CIFAR-10 ($\rho = 0.3$)	SVHN ($\rho = 0.3$)	CIFAR-100 ($\rho = 0.05$)
Full Model	90.35%	95.81%	68.53%
CLS-w/o-cld	53.86%	73.77%	26.62%
CLS-w/o-am	82.61%	92.92%	50.89%
CLS-w/o-ga	88.67%	94.34%	65.17%
CLS-w/o-am-ga	78.78%	91.36%	48.79%
CLS-w/o-w_1	87.83%	94.78%	58.99%
CLS-w/o-w_2	89.04%	94.80%	60.92%
CLS-w/o-$(w_1 + w_2)$	86.61%	93.53%	54.56%

Table 3: Ablation study on CIFAR-10 and SVHN with $\rho = 0.3$, and on CIFAR-100 with $\rho = 0.05$.

Conclusion

In this paper, we formalized a new and more practical learning scenario named as federated partial label learning (FedPLL) to investigate federated learning with low-quality partial labels. We proposed a novel method, FedPLL-LAAR, to address the entangled challenges of FL and PLL by designing effective label disambiguation, local-adaptive data augmentation, and local-global regularization mechanisms, which can effectively tackle the difficulties of FedPLL such as local data insufficiency, client drift, and noisy partial labels. Extensive experiments were conducted on multiple datasets, and the results validated the effectiveness of the proposed approach for federated partial label learning.

References

Acar, D. A. E.; Zhao, Y.; Navarro, R. M.; Mattina, M.; Whatmough, P. N.; and Saligrama, V. 2021. Federated learning based on dynamic regularization. *arXiv preprint arXiv:2111.04263*.

Chen, Y.; Yang, X.; Qin, X.; Yu, H.; Chen, B.; and Shen, Z. 2020. Focus: Dealing with label quality disparity in federated learning. *arXiv preprint arXiv:2001.11359*.

Cour, T.; Sapp, B.; and Taskar, B. 2011. Learning from partial labels. *Journal of Machine Learning Research*, 12: 1501–1536.

Feng, L.; and An, B. 2018. Leveraging Latent Label Distributions for Partial Label Learning. In *International Joint Conference on Artificial Intelligence (IJCAI)*.

Feng, L.; and An, B. 2019. Partial label learning with self-guided retraining. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Feng, L.; Lv, J.; Han, B.; Xu, M.; Niu, G.; Geng, X.; An, B.; and Sugiyama, M. 2020. Provably consistent partial-label learning. *Advances in Neural Information Processing Systems (NeurIPS)*.

Guo, P.; Wang, P.; Zhou, J.; Jiang, S.; and Patel, V. M. 2021. Multi-institutional collaborations for improving deep learning-based magnetic resonance image reconstruction using federated learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Han, S.; Park, S.; Wu, F.; Kim, S.; Wu, C.; Xie, X.; and Cha, M. 2022. FedX: Unsupervised Federated Learning with Cross Knowledge Distillation. In *European Conference on Computer Vision (ECCV)*.

Hsu, T.-M. H.; Qi, H.; and Brown, M. 2019. Measuring the effects of non-identical data distribution for federated visual classification. *arXiv preprint arXiv:1909.06335*.

Hüllermeier, E.; and Beringer, J. 2006. Learning from ambiguously labeled examples. *Intelligent Data Analysis*, 10(5): 419–439.

Jiang, L.; Huang, D.; Liu, M.; and Yang, W. 2020. Beyond synthetic noise: Deep learning on controlled noisy labels. In *International conference on machine learning (ICML)*.

Jin, R.; and Ghahramani, Z. 2002. Learning with multiple labels. In *Advances in neural information processing systems (NeurIPS)*.

Kairouz, P.; McMahan, H. B.; Avent, B.; Bellet, A.; Bennis, M.; Bhagoji, A. N.; Bonawitz, K.; Charles, Z.; Cormode, G.; Cummings, R.; et al. 2021. Advances and open problems in federated learning. *Foundations and Trends® in Machine Learning*, 14(1–2): 1–210.

Karimireddy, S. P.; Kale, S.; Mohri, M.; Reddi, S.; Stich, S.; and Suresh, A. T. 2020. Scaffold: Stochastic controlled averaging for federated learning. In *International Conference on Machine Learning (ICML)*.

Konečnỳ, J.; McMahan, H. B.; Yu, F. X.; Richtárik, P.; Suresh, A. T.; and Bacon, D. 2016. Federated learning: Strategies for improving communication efficiency. *NeurIPS Workshop on Private Multi-Party Machine Learn.*

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images. *Technical Report*.

Li, T.; Sahu, A. K.; Talwalkar, A.; and Smith, V. 2020a. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3): 50–60.

Li, T.; Sahu, A. K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; and Smith, V. 2020b. Federated optimization in heterogeneous networks. In *Conference on Machine Learning and Systems (MLSys)*.

Liu, L.; and Dietterich, T. 2012. A conditional multinomial mixture model for superset label learning. *Advances in neural information processing systems (NeurIPS)*.

Long, G.; Tan, Y.; Jiang, J.; and Zhang, C. 2020. Federated learning for open banking. In *Federated learning*, 240–254. Springer.

Lv, J.; Xu, M.; Feng, L.; Niu, G.; Geng, X.; and Sugiyama, M. 2020. Progressive identification of true labels for partial-label learning. In *International Conference on Machine Learning (ICML)*.

Lyu, G.; Feng, S.; Li, Y.; Jin, Y.; Dai, G.; and Lang, C. 2020. Hera: Partial label learning by combining heterogeneous loss with sparse and low-rank regularization. *ACM Transact. on Intellig. Systems and Technology*, 11(3): 1–19.

Lyu, G.; Feng, S.; Wang, T.; Lang, C.; and Li, Y. 2019. GM-PLL: graph matching based partial label learning. *IEEE Transact. on Knowledge and Data Engineering*, 521–535.

McMahan, B.; Moore, E.; Ramage, D.; Hampson, S.; and y Arcas, B. A. 2017. Communication-efficient learning of deep networks from decentralized data. In *Intern. Conference on Artificial Intelligence and Statistics (AISTATS)*.

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; and Ng, A. Y. 2011. Reading digits in natural images with unsupervised feature learning. In *NeurIPS Workshop on Deep Learning and Unsupervised Feature Learning*.

Nguyen, N.; and Caruana, R. 2008. Classification with partial labels. In *ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*.

Reddi, S.; Charles, Z.; Zaheer, M.; Garrett, Z.; Rush, K.; Konečnỳ, J.; Kumar, S.; and McMahan, H. B. 2020. Adaptive federated optimization. *arXiv preprint arXiv:2003.00295*.

Tuor, T.; Wang, S.; Ko, B. J.; Liu, C.; and Leung, K. K. 2021. Overcoming noisy and irrelevant data in federated learning. In *International Conference on Pattern Recognition (ICPR)*.

Wang, H.; Xiao, R.; Li, Y.; Feng, L.; Niu, G.; Chen, G.; and Zhao, J. 2022. PiCO: Contrastive Label Disambiguation for Partial Label Learning. In *International Conference on Learning Representations (ICLR)*.

Wen, H.; Cui, J.; Hang, H.; Liu, J.; Wang, Y.; and Lin, Z. 2021. Leveraged weighted loss for partial label learning. In *International Conference on Machine Learning (ICML)*.

Wu, D.-D.; Wang, D.-B.; and Zhang, M.-L. 2022. Revisiting Consistency Regularization for Deep Partial Label Learning. In *International Conference on Machine Learning (ICML)*.

Xu, J.; Chen, Z.; Quek, T. Q.; and Chong, K. F. E. 2022. FedCorr: Multi-Stage Federated Learning for Label Noise Correction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Xu, N.; Qiao, C.; Geng, X.; and Zhang, M.-L. 2021. Instance-Dependent Partial Label Learning. *Advances in Neural Information Processing Systems (NeurIPS)*.

Yan, Y.; and Guo, Y. 2020. Partial label learning with batch label correction. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Yang, Q.; Liu, Y.; Chen, T.; and Tong, Y. 2019. Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology*, 10(2): 1–19.

Yang, S.; Park, H.; Byun, J.; and Kim, C. 2022. Robust federated learning with noisy labels. *IEEE Intelligent Systems*, 37(2): 35–43.

Yao, Y.; Deng, J.; Chen, X.; Gong, C.; Wu, J.; and Yang, J. 2020. Deep discriminative cnn with temporal ensembling for ambiguously-labeled image classification. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Yu, F.; and Zhang, M.-L. 2016. Maximum margin partial label learning. In *Asian Conference on Machine Learning (ACML)*.

Zhang, F.; Feng, L.; Han, B.; Liu, T.; Niu, G.; Qin, T.; and Sugiyama, M. 2022. Exploiting Class Activation Value for Partial-Label Learning. In *International Conference on Learning Representations (ICLR)*.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2017. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*.

Zhang, M.-L.; and Yu, F. 2015. Solving the partial label learning problem: An instance-based approach. In *International Joint Conference on Artificial Intelligence (IJCAI)*.

Zhao, N.; Wu, Z.; Lau, R. W.; and Lin, S. 2020. What makes instance discrimination good for transfer learning? In *International Conference on Learning Representations (ICLR)*.