FedCorr: Multi-Stage Federated Learning for Label Noise Correction

Jingyi Xu[1*] Zihan Chen[1,2*] Tony Q.S. Quek[1] Kai Fong Ernest Chong[1†]
[1]Singapore University of Technology and Design [2]National University of Singapore
{jinyi_xu,zihan_chen}@mymail.sutd.edu.sg {tonyquek,ernest_chong}@sutd.edu.sg

CVPR 2022

Introduction

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

heterogeneous label noise

have data with label noise at different noise levels. Hence, the deployment of practical FL systems would face challenges brought by discrepancies in two aspects i): local data statistics [5, 12, 19, 24], and ii): local label quality [4, 35]. Although recent works explored the discrepancy in local data statistics in FL, and learning with label noise in centralized learning (CL), there is at present no unified approach for tackling both challenges simultaneously in FL.

privacy requirements

due to the limited sizes of local datasets. These CL methods cannot be applied on the global sever or across multiple clients due to FL privacy requirements. So, it is necessary and natural to adopt a more general framework that jointly considers the two discrepancies, for a better emulation of real-world data heterogeneity. Most importantly, privacy-preserving label correction should be incorporated in training to improve robustness to data heterogeneity in FL.

multi-stage FL framework：FedCorr

Introduction

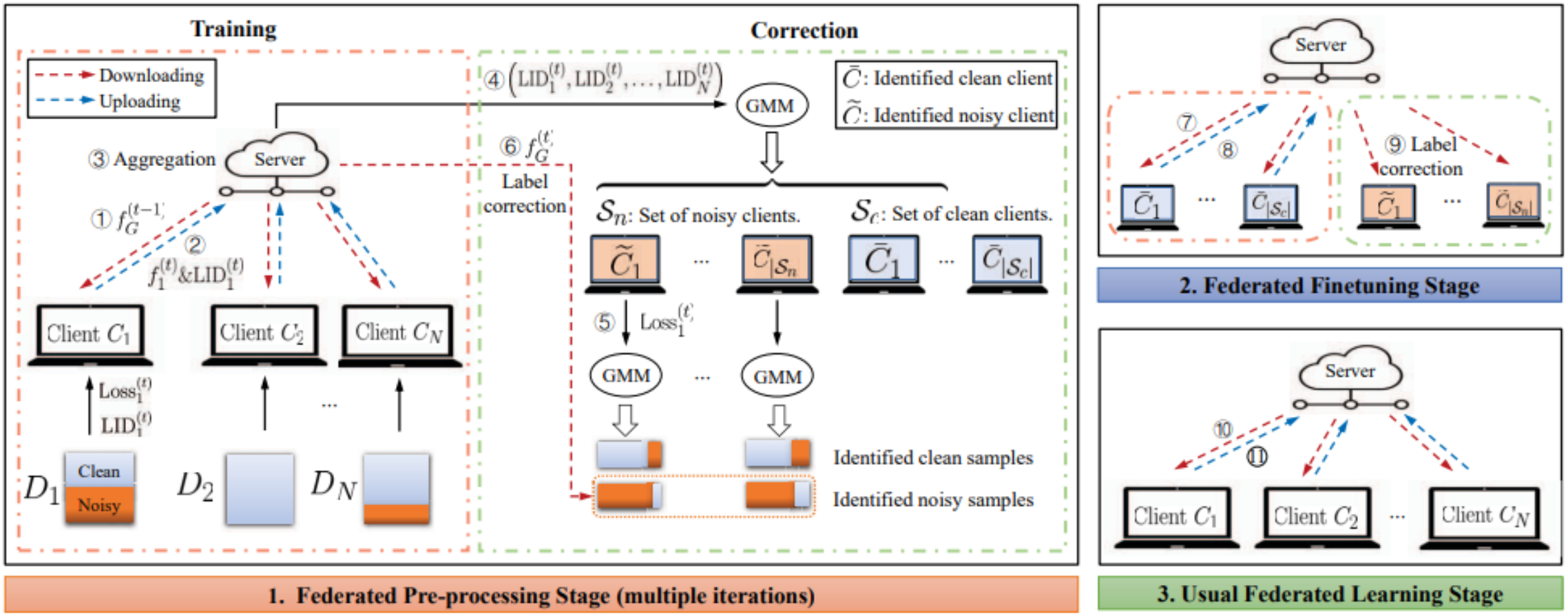


Figure 1. An overview of `FedCorr`, organized into three stages. Algorithm steps are numbered accordingly.

Introduction

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

Main contributions

- We propose a general multi-stage FL framework `FedCorr` to tackle data heterogeneity, with respect to both local label quality and local data statistics.
- We propose a general framework for easy generation of federated synthetic label noise and diverse (e.g. non-IID) client data partitions.

- We identify noisy clients via LID scores, and identify noisy labels via per-sample losses. We also propose an adaptive local proximal regularization term based on estimated local noise levels.
- We demonstrate that `FedCorr` outperforms state-of-the-art FL methods on multiple datasets with different noise levels, for both IID and non-IID data partitions.

Related Works

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

- **Federated methods**
- **FedProx/FedDyn/SCAFFOLD/PoC**
- **robust aggregation methods/reputation mechanism-based contribution examining/credibility-based re-weighting/distillation-based semisupervised learning**

tifying noisy labels. Even when these methods are used to detect noisy clients, either there is no mechanism for further label correction at the noisy clients [7, 17, 28, 33], or the effect of noisy labels is mitigated with the aid of an auxiliary dataset, without any direct label correction [4, 13].

•Local intrinsic dimension (LID)

Informally, LID [10] is a measure of the intrinsic dimensionality of the data manifold. In comparison to other measures, LID has the potential for wider applications as it makes no further assumptions on the data distribution beyond continuity. The key underlying idea is that at each dat-

factor of r. Specifically, when we have two m-dimensional Euclidean balls with volumes V_1, V_2, and with radii r_1, r_2, we can compute m as follows:

$$\frac{V_2}{V_1} = \left(\frac{r_2}{r_1}\right)^m \Rightarrow m = \frac{\log(V_2/V_1)}{\log(r_2/r_1)}. \qquad (1)$$

Related Works

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

•Local intrinsic dimension (LID)

We shall now formally define LID. Suppose we have a dataset consisting of vectors in $\mathbb{R}^n$. We shall treat this dataset as samples drawn from an n-variate distribution $\overline{\mathcal{D}}$. For any $x \in \mathbb{R}^n$, let Y_x be the random variable representing the (non-negative) distance from x to a randomly selected point y drawn from $\overline{\mathcal{D}}$, and let $F_{Y_x}(t)$ be the cumulative distribution function of Y_x. Given $r > 0$ and a sample point x drawn from $\overline{\mathcal{D}}$, define the *LID of x at distance r* to be

$$\mathrm{LID}_x(r) := \lim_{\varepsilon \to 0} \frac{\log F_{Y_x}((1+\varepsilon)r) - \log F_{Y_x}(r)}{\log(1+\varepsilon)},$$

$$\text{limit } \mathrm{LID}_x = \lim_{r \to 0} \mathrm{LID}_x(r).$$

包含x的光滑流形函数的近似值

$$\widehat{\mathrm{LID}}(x) = -\left(\frac{1}{k}\sum_{i=1}^{k} \log \frac{r_i(x)}{r_{max}(x)}\right)^{-1}, \qquad (2)$$

where $r_i(x)$ denotes the distance between x and its i-th nearest neighbor, and $r_{max}(x)$ is the maximum distance from x among the k nearest neighbors.

Method

- preliminaries
- pre-processing
- finetuning
- usual training

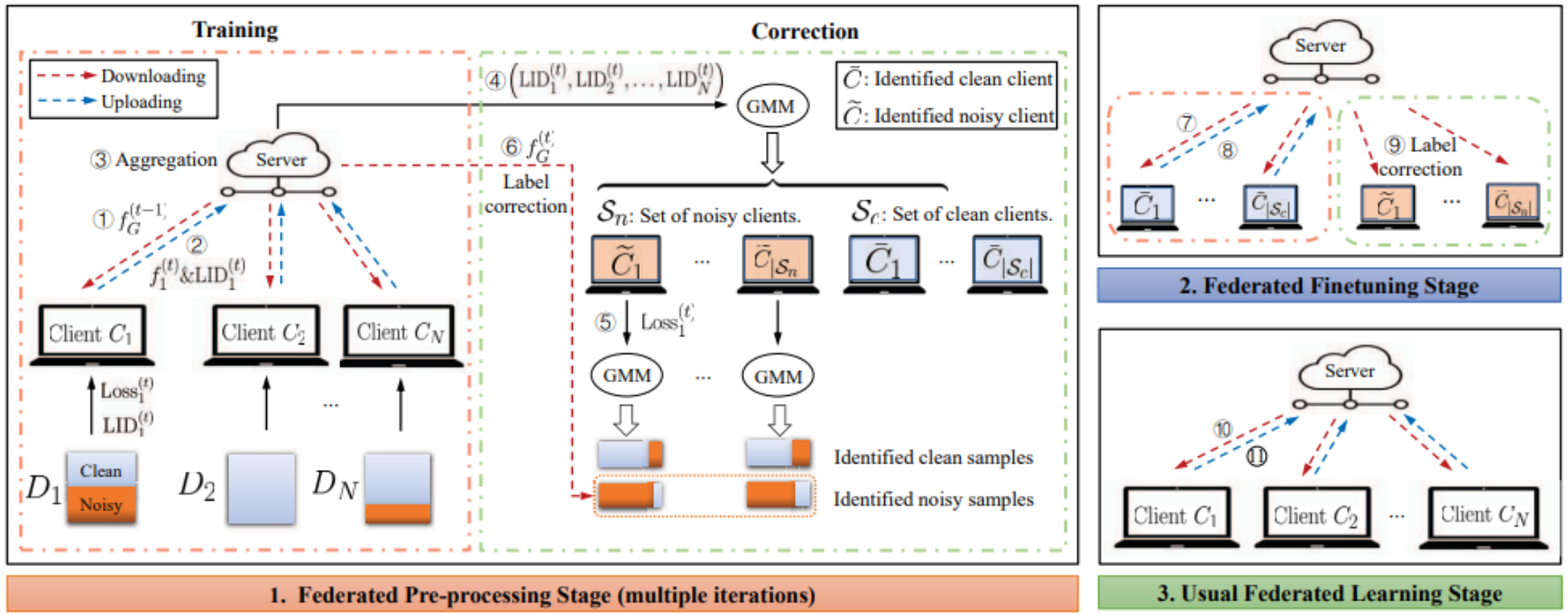


Figure 1. An overview of FedCorr, organized into three stages. Algorithm steps are numbered accordingly.

Method

- **preliminaries**

Consider an FL system with N clients and an M-class dataset $\mathcal{D} = \{\mathcal{D}_k\}_{k=1}^{N}$, where each $\mathcal{D}_k = \{(x_k^i, y_k^i)\}_{i=1}^{n_k}$ denotes the local dataset for client k. Let $\mathcal{S}$ denote the set of all N clients, and let $w_k^{(t)}$ (resp. $w^{(t)}$) denote the local model weights of client k (resp. global model weights obtained by aggregation) at the end of communication round t. At the end of round t, the global model $f_G^{(t)}$ would have its weights $w^{(t)}$ updated as follows:

$$w^{(t)} \leftarrow \sum_{k \in \mathcal{S}_t} \frac{|\mathcal{D}_k|}{\sum_{i \in \mathcal{S}_t} |\mathcal{D}_i|} w_k^{(t)}, \tag{3}$$

Method

- **preliminaries**
- **IID partition**

  unformly distributied
- **non-IID partition**

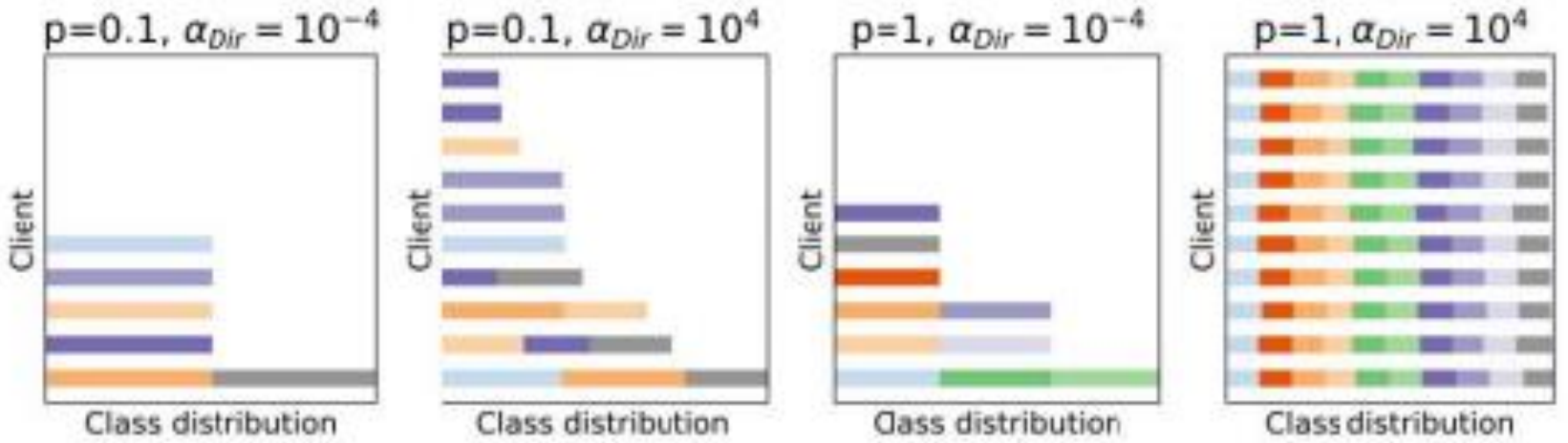


Figure 2. Depiction of non-IID partitions for different parameters.

Method

- **preliminaries**
- **noise level**

$$\mu_k = \begin{cases} u \sim U(\tau, 1), & \text{with probability } \rho; \\ 0, & \text{with probability } 1 - \rho. \end{cases} \tag{4}$$

When $\mu_k \neq 0$, the $100 \cdot \mu_k\%$ noisy samples are chosen uniformly at random, and are assigned random labels, selected uniformly from the M classes.

Method

- **preliminaries**
- **LID scores**

Experiments have shown that given the same training process, models trained on a dataset with label noise tend to have larger LID scores as compared to models trained on the same dataset with clean labels [22, 23]. Intuitively,

Method

- **pre-precessing**

- All clients will participate in each iteration. Clients are selected without replacement, using a small fraction.
- An adaptive local proximal term is added to the loss function, and mixup data augmentation is used.
- Each client computes its LID score and per-sample cross-entropy loss after local training and sends its LID score together with local model updates to the server.

Method

- **pre-precessing**
- **Client iteration and fraction scheduling**

The pre-processing stage is divided into T_1 iterations. In each iteration, every client participates exactly once. Every iteration is organized by communication rounds, similar to the usual FL, but with two key differences: a small fraction is used, and clients are selected without replacement. Each iteration ends when all clients have participated.

Method

- **pre-precessing**
- **Mixup and local proximal regularization**

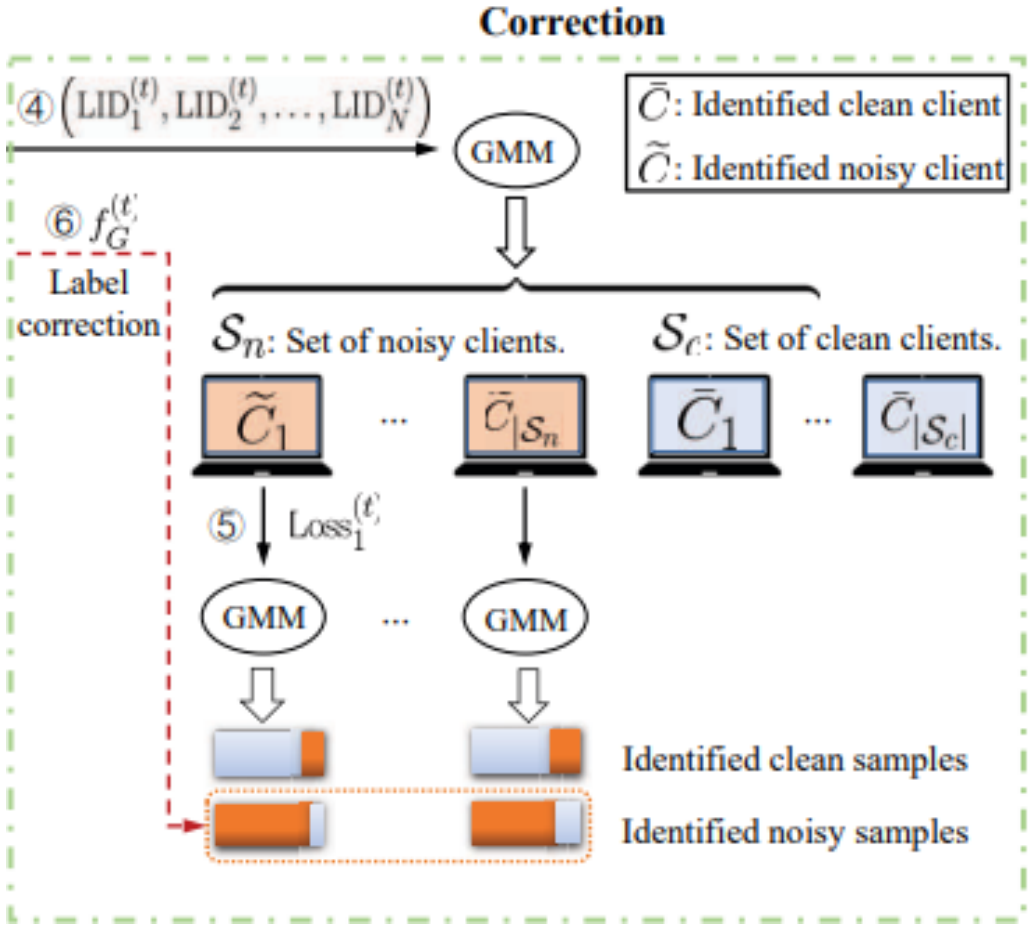
$$L(X_b) = L_{CE}\left(f_k^{(t)}(\tilde{X}_b), \tilde{Y}_b\right) + \beta\hat{\mu}_k^{(t-1)}\left\|w_k^{(t)} - w^{(t-1)}\right\|^2. \quad (5)$$

Here, $f_k^{(t)} = f(\cdot; w_k^{(t)})$ denotes the local model of client k in round t, and $w^{(t-1)}$ denotes the weights of the global model obtained in the previous round $t-1$. The first term in (5) represents the cross-entropy loss on the mixup augmentation of (X_b, Y_b), while the second term in (5) is an adaptive local proximal regularization term, where $\hat{\mu}_k^{(t-1)}$ is the estimated noise level of client k to be defined later. It should be noted that our local proximal regularization term is only applied in the pre-processing stage.

[3,16]. Mixup generates new samples $(\tilde{x}, \tilde{y})$ as convex combinations of randomly selected pairs of samples (x_i, y_i) and (x_j, y_j), given by $\tilde{x} = \lambda x_i + (1-\lambda)x_j$, $\tilde{y} = \lambda y_i + (1-\lambda)y_j$, where $\lambda \sim \text{Beta}(\alpha, \alpha)$, and $\alpha \in (0, \infty)$. (We use $\alpha = 1$ in

Method

- **pre-precessing**
- **Identification of noisy clients and noisy samples**

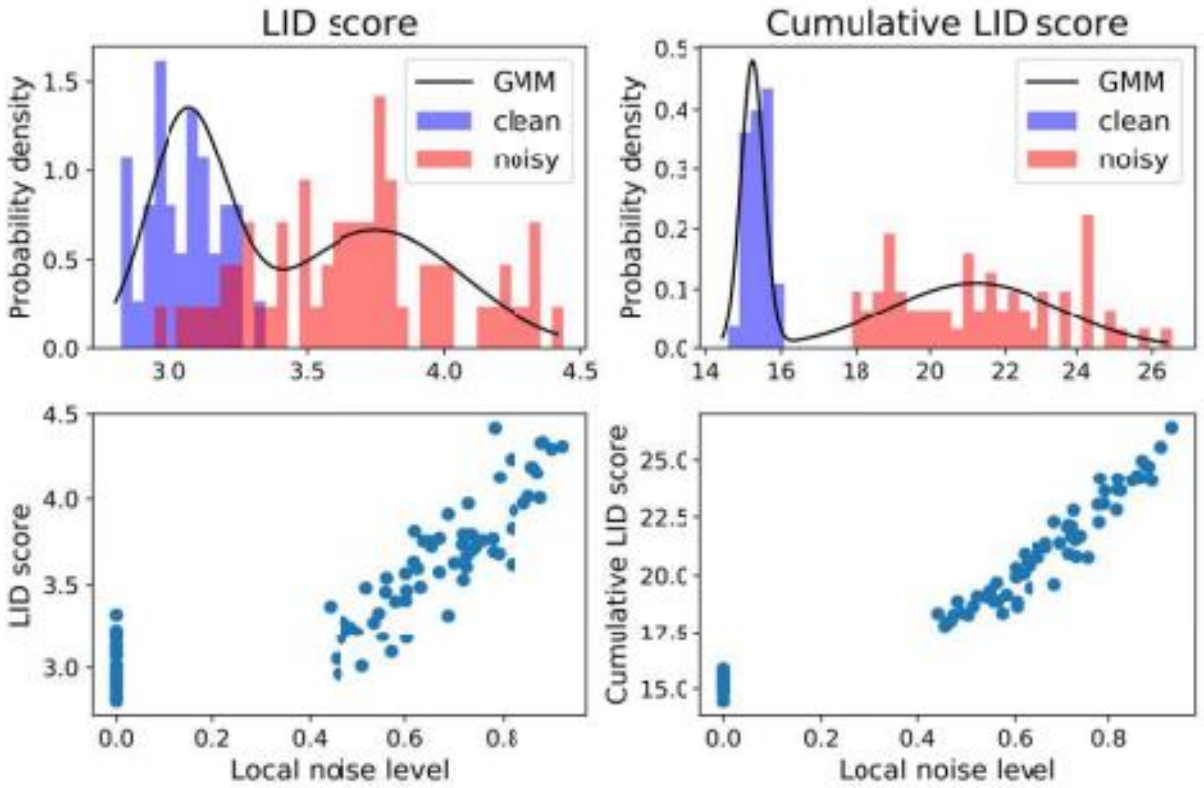


$$\widetilde{\mathcal{D}}_k^n = \underset{\substack{\tilde{\mathcal{D}} \subseteq \mathcal{D}_k^n \\ |\tilde{\mathcal{D}}| = \pi \cdot |\mathcal{D}_k^n|}}{\arg\max} L_{CE}(\tilde{\mathcal{D}}; f_G^{(t)}); \tag{6}$$

$$\widetilde{\mathcal{D}}_k^{n\prime} = \left\{(x, y) \in \widetilde{\mathcal{D}}_k^n \middle| \max(f_G^{(t)}(x)) \geq \theta\right\}; \tag{7}$$

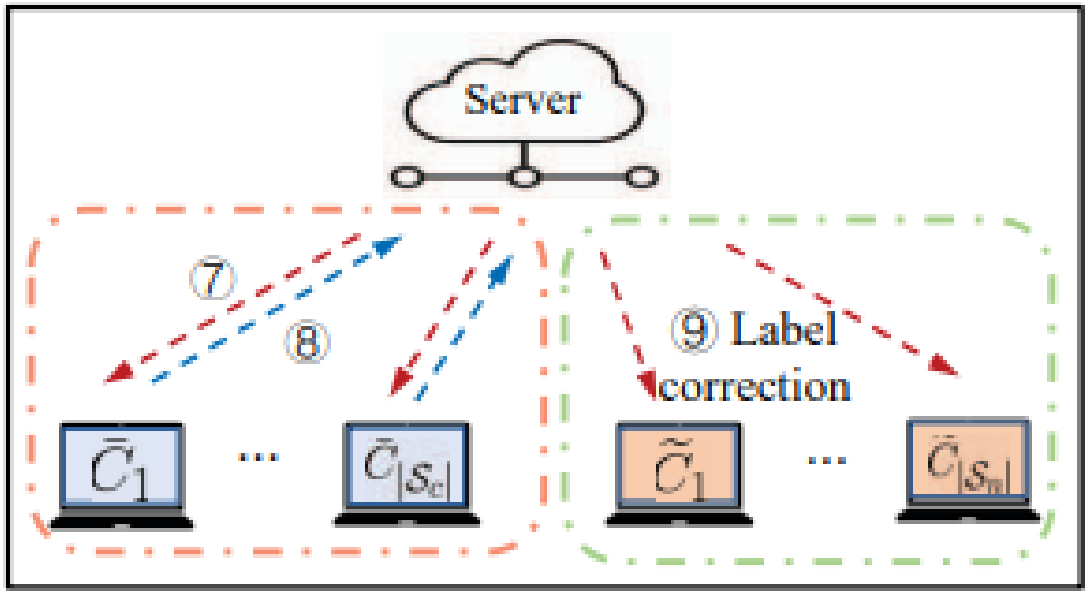
Method

- **pre-precessing**
- **Identification of noisy clients and noisy samples**



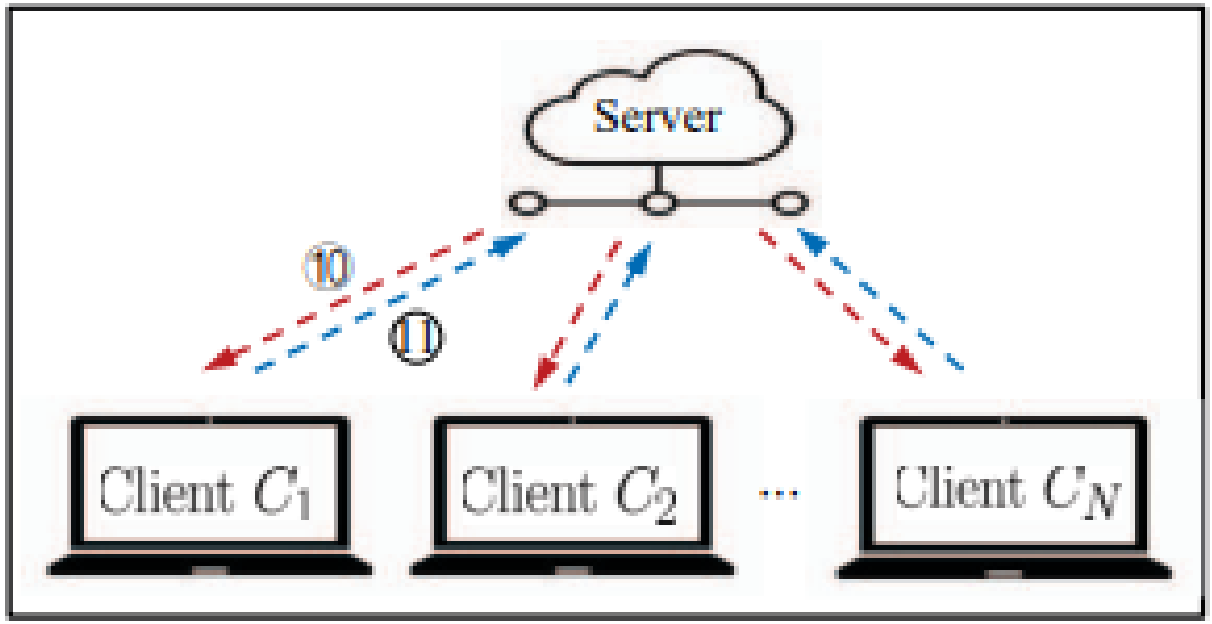
Method

- **Federated finetuning stage**



2. Federated Finetuning Stage

- **Federated usual training stage**



3. Usual Federated Learning Stage

Method

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

// Federated Pre-processing Stage

$(\hat{\mu}_1^{(0)}, \ldots, \hat{\mu}_N^{(0)}) \leftarrow (0, \ldots, 0)$ *// estimated noise levels*

for $t = 1$ **to** T_1 **do**

  $\mathcal{S}$ =Shuffle($\{1, \ldots, N\}$)

  $w_{inter} \leftarrow w^{(t-1)}$ *// intermediary weights*

  for $k \in \mathcal{S}$ **do**

    $w_k^{(t)} \leftarrow$ weights that minimize loss function (5)

    Upload weights $w_k^{(t)}$ and LID score to server

  Update global model $w^{(t)} \leftarrow w_{inter}$

  Divide all clients into clean set $\mathcal{S}_c$ and noisy set $\mathcal{S}_n$ based on cumulative LID scores via GMM

  for noisy client $k \in \mathcal{S}_n$ **do**

    Divide $\mathcal{D}_k$ into clean subset $\mathcal{D}_k^c$ and noisy subset $\mathcal{D}_k^n$ based on per-sample losses via GMM

    $\hat{\mu}_k^{(t)} \leftarrow \frac{|\mathcal{D}_k^n|}{|\mathcal{D}_k|}$ *// update estimated noise level*

    $y_k^{(i)} \leftarrow \arg\max f(x_k^{(i)}; w^{(i)}), \forall (x_k^{(i)}, y_k^{(i)}) \in \mathcal{D}_k^n$

// Federated Finetuning Stage

$\mathcal{S}_c \leftarrow \{k | k \in \mathcal{S}, \mu_k < 0.1\}, \mathcal{S}_n \leftarrow \mathcal{S} \setminus \mathcal{S}_c$.

for $t = T_1 + 1$ **to** $T_1 + T_2$ **do**

  Update $w_k^{(t)}$ by usual `FedAvg` among clients in $\mathcal{S}_c$

for Noisy client $k \in \mathcal{S}_n$ **do**

  $y_k^{(i)} \leftarrow \arg\max f(x_k^{(i)}; w^{(i)}), \forall (x_k^{(i)}, y_k^{(i)}) \in \mathcal{D}_k$

// Usual Federated Learning Stage

for $t = T_1 + T_2 + 1$ **to** $T_1 + T_2 + T_3$ **do**

  Update $w_k^{(t)}$ by usual `FedAvg` among all clients

return $f_G^{\text{final}} := f(\cdot; w^{(T_1+T_2+T_3)})$

Experiments

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

datasets

Dataset	CIFAR-10	CIFAR-100	Clothing1M
Size of $\mathcal{D}_{train}$	50,000	50,000	1,000,000
# of classes	10	100	14
# of clients	100	50	500
Fraction γ	0.1	0.1	0.02
Architecture	ResNet-18	ResNet-34	pre-trained ResNet-50

Table 1. List of datasets used in our experiments.

FedCorr: robust to discrepancies in both data statistics and label quality

Setting	Method	Best Test Accuracy (%) ± Standard Deviation (%)						
		$\rho = 0.0$	$\rho = 0.4$		$\rho = 0.6$		$\rho = 0.8$	
		$\tau = 0.0$	$\tau = 0.0$	$\tau = 0.5$	$\tau = 0.0$	$\tau = 0.5$	$\tau = 0.0$	$\tau = 0.5$
Centralized (for reference)	JointOpt	93.73±0.21	92.29±0.37	92.11±0.21	91.26±0.46	88.42±0.33	89.18±0.29	85.62±1.17
	DivideMix	95.64±0.05	96.39±0.09	96.17±0.05	96.07±0.06	94.59±0.09	94.21±0.27	94.36±0.16
Federated	FedAvg	93.11±0.12	89.46±0.39	88.31±0.80	86.09±0.50	81.22±1.72	82.91±1.35	72.00±2.76
	FedProx	92.28±0.14	88.54±0.33	88.20±0.63	85.80±0.41	85.25±1.02	84.17±0.77	80.59±1.49
	RoFL	88.33±0.07	88.25±0.33	87.20±0.26	87.77±0.83	83.40±1.20	87.08±0.65	74.13±3.90
	ARFL	92.76±0.08	85.87±1.85	83.14±3.45	76.77±1.90	64.31±3.73	73.22±1.48	53.23±1.67
	JointOpt	88.16±0.18	84.42±0.70	83.01±0.88	80.82±1.19	74.09±1.43	76.13±1.15	66.16±1.71
	DivideMix	77.96±0.15	77.35±0.20	74.40±2.69	72.67±3.39	72.83±0.30	68.66±0.51	68.04±1.38
	Ours	**93.82±0.41**	**94.01±0.22**	**94.15±0.18**	**92.93±0.25**	**92.50±0.28**	**91.52±0.50**	**90.59±0.70**

Table 2. Average (5 trials) and standard deviation of the best test accuracies of various methods on CIFAR-10 with IID setting at different noise levels (ρ: ratio of noisy clients, τ: lower bound of client noise level). The highest accuracy for each noise level is boldfaced.

FedCorr: robust to discrepancies in both data statistics and label quality

Method	Best Test Accuracy (%) ± Standard Deviation(%)			
	$\rho = 0.0$ $\tau = 0.0$	$\rho = 0.4$ $\tau = 0.5$	$\rho = 0.6$ $\tau = 0.5$	$\rho = 0.8$ $\tau = 0.5$
JointOpt (CL)	72.94±0.43	65.87±1.50	60.55±0.64	59.79±2.45
DivideMix (CL)	75.58±0.14	75.43±0.34	72.26±0.58	71.02±0.65
FedAvg	72.41±0.18	64.41±1.79	53.51±2.85	44.45±2.86
FedProx	71.93±0.13	65.09±1.46	57.51±2.01	51.24±1.60
RoFL	67.89±0.65	59.42±2.69	46.24±3.59	36.65±3.36
ARFL	72.05±0.28	51.53±4.38	33.03±1.81	27.47±1.08
JointOpt	67.49±0.36	58.43±1.88	44.54±2.87	35.25±3.02
DivideMix	45.91±0.27	43.25±1.01	40.72±1.41	38.91±1.25
Ours	**72.56±2.07**	**74.43±0.72**	**66.78±4.65**	**59.10±5.12**

Table 3. Average (5 trials) and standard deviation of the best test accuracies on CIFAR-100 with IID setting.

FedCorr: robust to discrepancies in both data statistics and label quality

Method\(p, α_{Dir})	$(0.7, 10)$	$(0.7, 1)$	$(0.3, 10)$
FedAvg	78.88±2.34	75.98±2.92	67.75±4.38
FedProx	83.32±0.98	80.40±0.94	73.86±2.41
RoFL	79.56±1.39	72.75±2.21	60.72±3.23
ARFL	60.19±3.33	55.86±3.30	45.78±2.84
JointOpt	72.19±1.59	66.92±1.89	58.08±2.18
DivideMix	65.70±0.35	61.68±0.56	56.67±1.73
Ours	**90.52±0.89**	**88.03±1.08**	**81.57±3.68**

Table 4. Average (5 trials) and standard deviation of the best test accuracies of different methods on CIFAR-10 with different non-IID setting. The noise level is $(\rho, \tau) = (0.6, 0.5)$.

FedCorr: robust to discrepancies in both data statistics and label quality

Settings	FedAvg	FedProx	RoFL	ARFL	JointOpt	Dividemix	Ours
FL	70.49	71.35	70.39	70.91	71.78	68.83	**72.55**
CL	-	-	-	-	72.23	74.76	-

Table 5. Best test accuracies on Clothing1M with non-IID setting. CL results are the accuracies reported in corresponding papers.

FedCorr: versatility

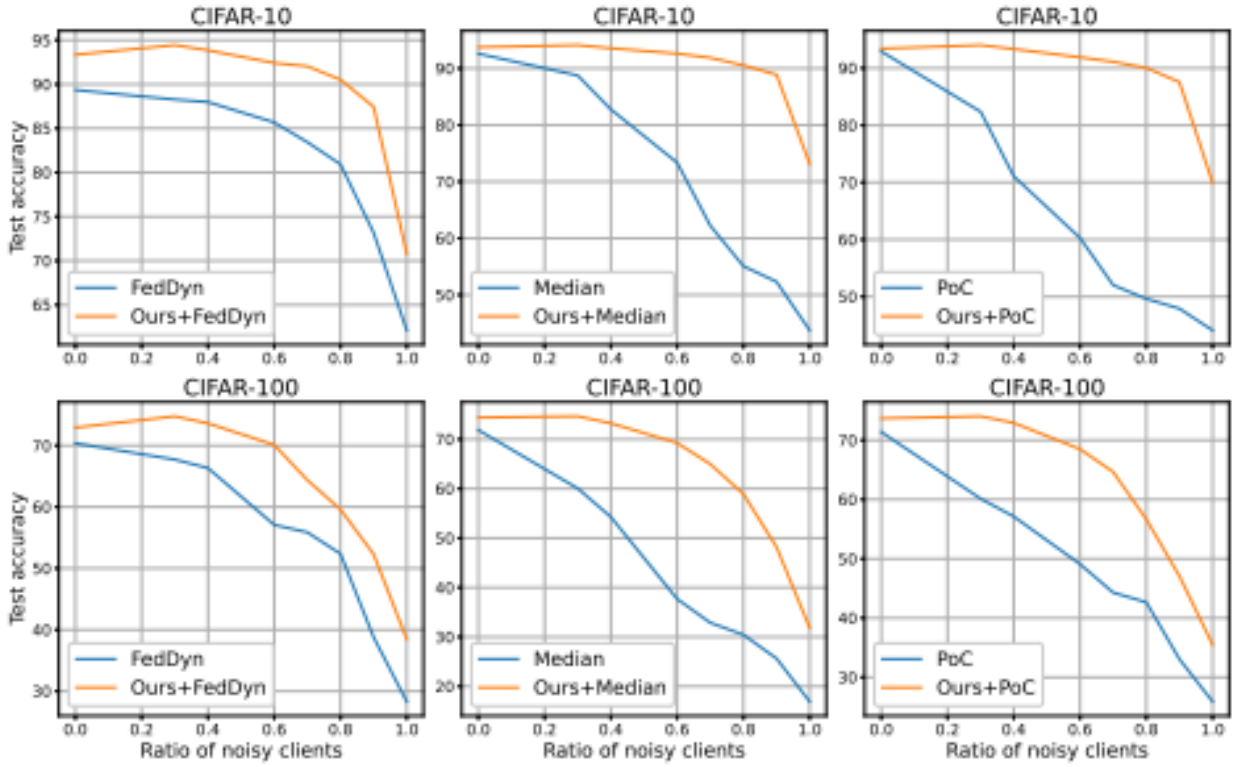


Figure 4. Best test accuracies of three FL methods combined with FedCorr on CIFAR-10/100 with multiple ρ and fixed $\tau = 0.5$.

Ablation study

Method	Best Test Accuracy (%) ± Standard Deviation (%)						
	$\rho = 0.0$	$\rho = 0.4$		$\rho = 0.6$		$\rho = 0.8$	
	$\tau = 0.0$	$\tau = 0.0$	$\tau = 0.5$	$\tau = 0.0$	$\tau = 0.5$	$\tau = 0.0$	$\tau = 0.5$
Ours	**93.82±0.41**	**94.01±0.22**	**94.15±0.18**	**92.93±0.25**	**92.50±0.28**	**91.52±0.50**	**90.59±0.70**
Ours w/o correction	92.85±0.66	93.71±0.20	93.60±0.21	92.15±0.29	91.77±0.65	90.48±0.56	88.77±1.10
Ours w/o frac. scheduling	86.05±1.47	85.59±1.10	78.44±7.90	80.29±2.62	77.96±3.65	76.67±3.48	72.71±5.03
Ours w/o local proximal	93.37±0.05	93.64±0.15	93.46±0.17	92.34±0.14	91.74±0.47	90.45±0.94	88.74±1.72
Ours w/o finetuning	92.71±0.18	93.06±0.15	92.62±0.28	91.41±0.14	89.31±0.90	89.62±0.40	83.81±2.59
Ours w/o usual training	93.11±0.10	93.53±0.17	93.46±0.14	92.16±0.24	91.50±0.51	90.62±0.59	88.97±1.37
Ours w/o mixup	90.63±0.70	88.83±1.88	91.34±0.39	87.79±0.89	87.50±1.33	87.86±0.53	83.29±1.78

Table 6. Ablation study results (average and standard deviation of 5 trials) on CIFAR-10.

Conclusion

FedCorr

- **tackle both local label quality and data statistic**
- **privacy-preserving label correction**
- **robustness**
- **not consider dynamic participation**
- **cumulative LID scores**

Thanks