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

Class Balanced Adaptive Pseudo Labeling for Federated Semi-Supervised Learning

Ming Li Qingli Li Yan Wang*
Shanghai Key Laboratory of Multidimensional Information Processing, East China Normal University
lm1640362161@gmail.com, qlli@cs.ecnu.edu.cn, ywang@cee.ecnu.edu.cn

CVPR 2023

Introduction

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

• Broadly three lines of FSSL methods

1) The first two lines consider that there are only limited labeled data in the central server or each client has partially labeled data.

2) The third line assumes that few clients have fully labeled data and the training datasets in other clients are fully unlabeled.(mainly focuses on)

•Main difficulties to train a third line FSSL model

1) There are no labeled data in unlabeled clients. Thus, the training can be easily biased without label guidance.

2) Due to Non-IID data, inaccurate supervisory signals may be generated in unlabeled clients via employing the model trained in labeled clients by either pseudo labeling or consistency regularization framework.

3) Due to the catastrophic forgetting problems in CNNs, with the training process of unlabeled clients going on, models may forget the knowledge learned on labeled clients and so decrease the prediction accuracy drastically.

Introduction

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

• CBAFed——Concretely, we present Class Balanced Adaptive Pseudo Labeling, namely CBAFed, by rethinking standard pseudolabeling methods in SSL.

• To handle the catastrophic forgetting problem, we propose a fixed pseudo labeling strategy, which builds a fixed set by letting pass informative unlabeled data and their pseudo labels at the beginning of the unlabeled client training.

• Due to the Non-IID and heterogeneous data partition problems in FL, training distribution of unlabeled data can be highly imbalanced, so existing thresholds are not suitable in FSSL. We design class balanced adaptive thresholds via considering the empirical distribution of all training data in local clients at the previous communication round.

• To enhance the learning ability and discover unlabeled data from tail classes, we propose to leverage information from so-called "not informative" unlabeled data.

• We introduce a residual weight connection method, to improve the robustness of the models in labeled clients and the central server, which skip connects weights from previous epoch or communication round to finally reach better optimum.

Related Works

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

- **FL**

• Vision Transformers(ViT) :self-attention-based architectures are more robust to distribution shifts and can converge tobetter optimum over heterogeneous data

- **SSL**

•Only consider pre-defined fixed threshold for pseudo labeling.
•While these methods perform well in centralized SSL, they all update pseudo labels after every batch's update of the model, which is not suitable in FSSL as shown in later section.

- **FSSL**

• Fed-Consist and FedIRM: do not consider data heterogeneity in federated learning
• RSCFed: perform random sub-sampling to reach consensus over clients.
It uses standard consistency regularization for unlabeled data, which still suffers from the Non-IID setting.

Method

Step 1) Warm up stage: train fully supervised models on only labeled clients using residual weight connection in a normal federated learning manner.

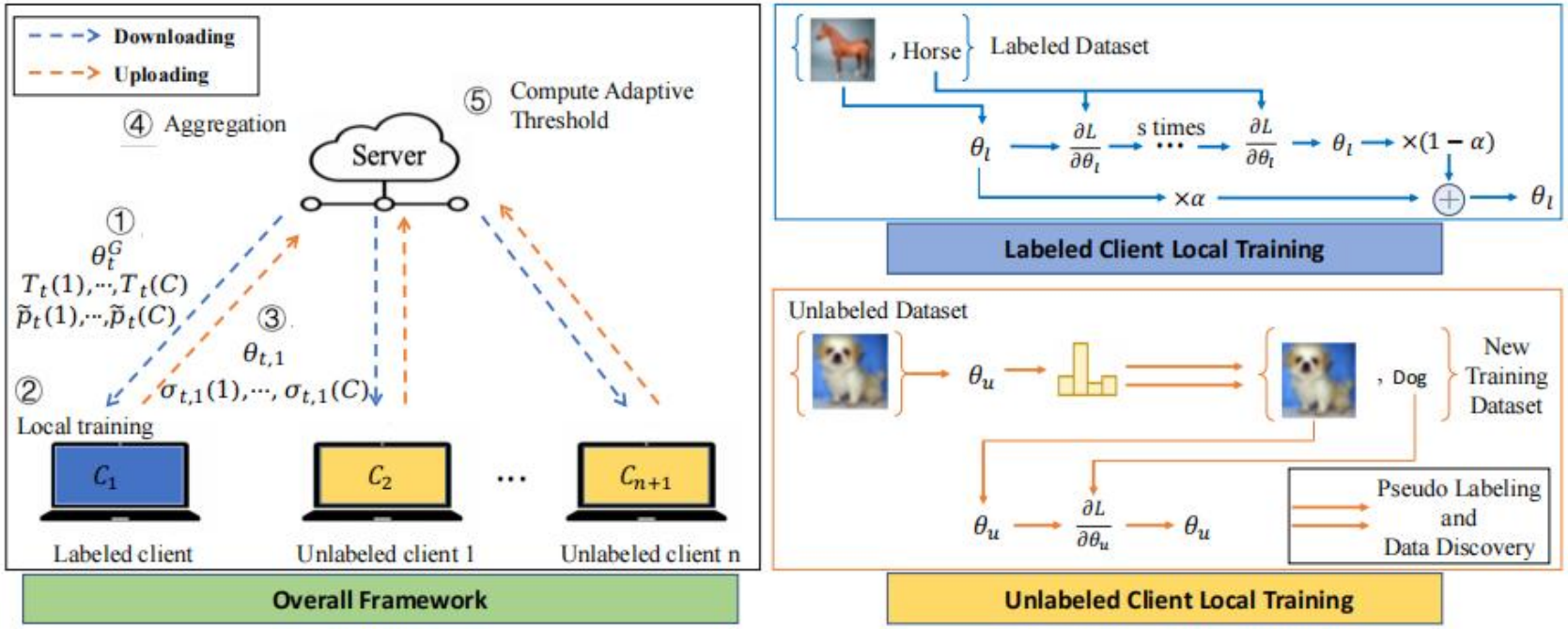
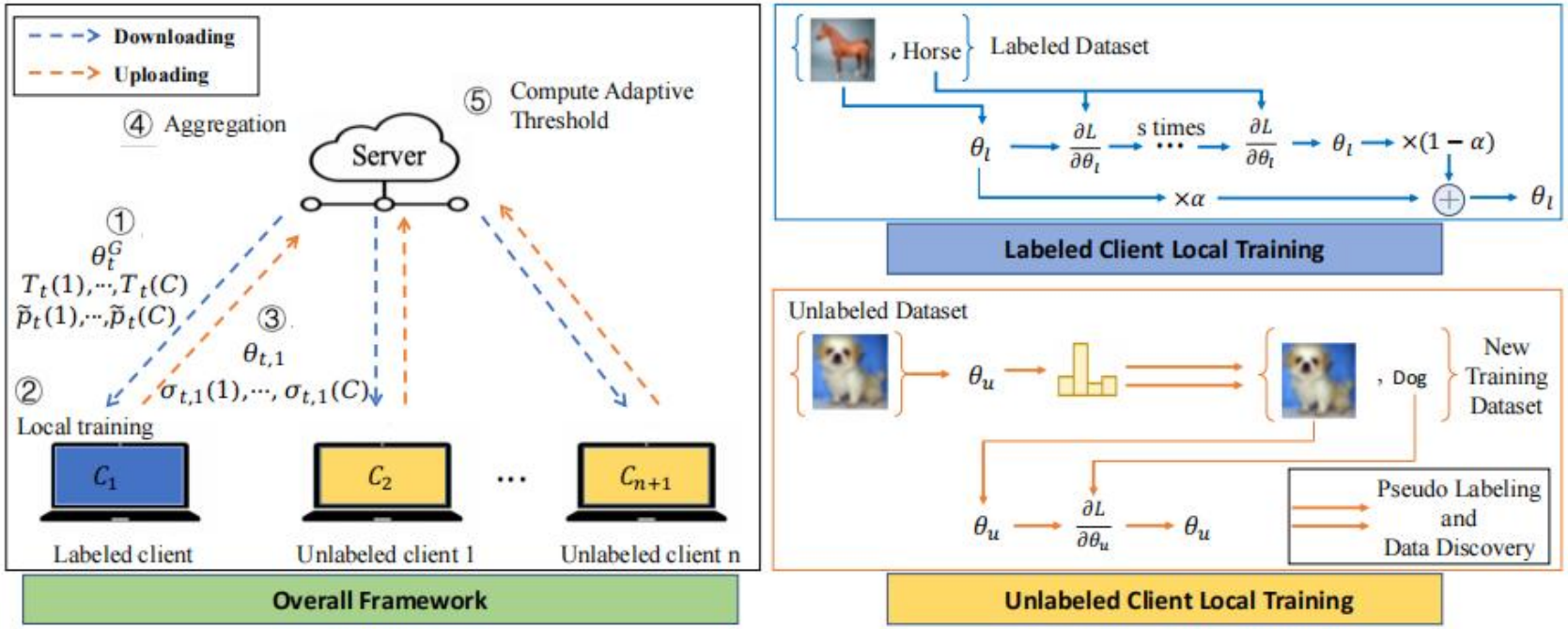


Figure 1. An overview of our CBAFed. In the central server (left side), the global model is aggregated with the returned local models (step ④) and the adaptive thresholds are calculated by the returned training data statistics (step ⑤). Then central server passes the global model, adaptive thresholds and class distribution to all local clients (step ①). After downloading these data, local clients perform local training on the right side (step ②). Labeled clients use labeled data to train the model with residual weight connection. Unlabeled clients obtain the new training dataset by adaptive pseudo labeling and tail class data discovery and use it to train the model. After local training, local clients return trained models and number of data in each class back to central server (step ③).

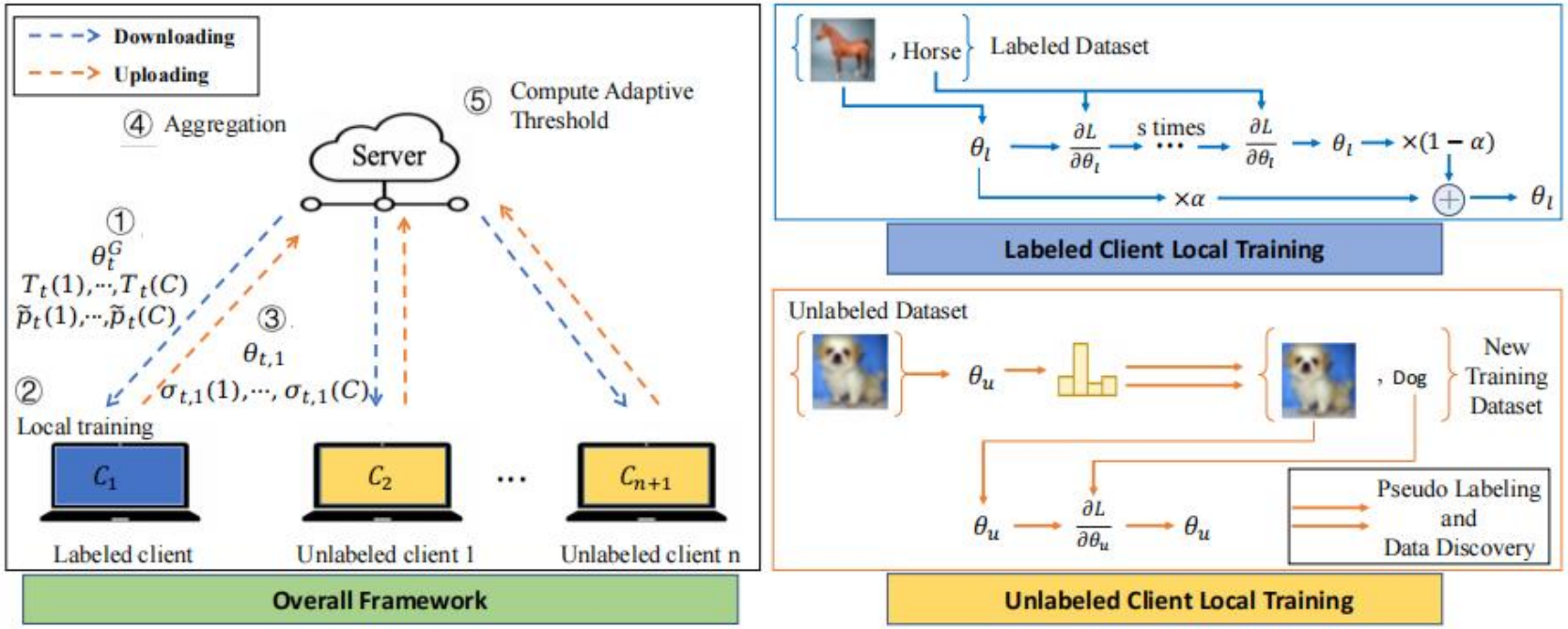
Method

Step 2) The central server computes the empirical class distribution and obtains the class balanced adaptive thresholds, then passes them to local clients.
Step 3) All local clients update local models, adaptive threshold and class distribution.
Labeled clients: train local models on all the data using proposed residual weight connection.
Unlabeled clients: acquire the fixed training set by the threshold and the tail class datasets, and train local models on the newly obtained training dataset.



Method

Step 4) The central server aggregates a new model with residual weight connection, computes the class distribution, and obtains the class balanced adaptive threshold. Then, the central server passes them to local clients.
Step 5) Repeat step (3)-(4) until the specified number of communication round is reached.



Method

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

• Residual Weight Connection

In ResNet, there is a skip connection between every layer.

There is a skip connection of model's parameters between training epochs (or communication rounds).

$$\theta^E = \begin{cases} \theta^E & E\%s \neq 0 \\ \alpha_1 \theta^{E-s} + (1-\alpha_1)\theta^E & E\%s = 0 \end{cases}$$

→ **Averaging model weights over training steps tends to produce a more accurate model than using the final weights directly.**

• Pseudo Labeling Methods

warm up:

$$\mathcal{L}_\ell = \frac{1}{|D_\ell|} \sum_{(X^\ell, y^\ell) \in D_\ell} H(y^\ell, p_m(y|\theta_t^\ell(X^\ell))),$$

$$\theta_{t+1}^G = \begin{cases} \sum_{\ell=1}^m \frac{|D_\ell|}{\sum_{i=1}^m |D_i|} \theta_t^\ell & t\%s \neq 0 \\ \alpha_1 \theta_{t+1-s}^G + (1-\alpha_1) \sum_{\ell=1}^m \frac{|D_\ell|}{\sum_{i=1}^m |D_i|} \theta_t^\ell & t\%s = 0 \end{cases} \quad (3)$$

fixed pseudo labeling:

$$\hat{y}_i^\mu = \arg\max p_m(y|\theta_t^\mu(X_i^\mu), i = 1, 2, \cdots, N_\mu.$$

$$\widetilde{\mathcal{D}}_\mu = \{(X_i^\mu, \hat{y}_i^\mu) \mid X_i^\mu \in D_\mu \wedge \max(p_m(y|\theta_t^\mu(X_i^\mu))) > \tau\}_{i=1}^{N_\mu}. \quad (5)$$

$$\mathcal{L}_\mu = \frac{1}{|\widetilde{\mathcal{D}}_\mu|} \sum_{(X^\mu, \hat{y}^\mu) \in \widetilde{\mathcal{D}}_\mu} H(\hat{y}^\mu, p_m(y|\theta_t^\mu(X^\mu))).$$

Method

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

- **Class Balanced Adaptive Threshold for Pseudo Labeling（CBAPL）**

Setting a fixed threshold usually makes the model fail to consider different learning status and learning difficulties of different classes.

- **Curriculum Pseudo Labeling**

$$\mathcal{T}_t(c) = \beta_t(c) \cdot \tau$$

Due to the Non-IID partition, the labeled data are not balanced, so purely using the number of selected unlabeled data to design threshold is improper.

→ **Introduce many noisy labels into training**

- **CBAPL**

$$\sigma_t^{\mu}(c) = \sum_{i=1}^{N_\mu} \mathbf{1}(\max(p_m(y|\theta_t^{\mu}(X_i^{\mu}))) > \mathcal{T}_t(c)) \mathbf{1}(\hat{y}_i^{\mu} = c)$$

$$\sigma_t^{\ell}(c) = \sum_{i=1}^{N_\ell} \mathbf{1}(y_i^{\ell} = c).$$

$$\sigma_t(c) = \sum_{\ell=1}^{m} \sigma_t^{\ell}(c) + \sum_{\mu=m+1}^{n+m} \sigma_t^{\mu}(c).$$

Method

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

• Class Balanced Adaptive Threshold for Pseudo Labeling（CBAPL）

empirical distribution

$$\widetilde{p}_t(c) = \frac{\sigma_t(c)}{\sum_{i=1}^{C} \sigma_t(i)}.$$

upper bound of threshold

$$\mathcal{T}_{t+1}(c) = \begin{cases} \tau_{t,c}, & \tau_{t,c} < \tau_h \\ \tau_h, & \tau_{t,c} \geq \tau_h \end{cases}$$

standard deviation

$$std(\widetilde{p}_t) = \sqrt{\frac{1}{C-1}\sum_{c=1}^{C}(\widetilde{p}_t(c) - \overline{p}_t)^2},$$

$$\overline{p}_t = \frac{1}{C}\sum_{c=1}^{C}\widetilde{p}_t(c).$$

fixed pseudo label training dataset

$$\widetilde{\mathcal{D}}_{t+1,\mu} = \{(X_i^\mu, \hat{y}_i^\mu) | X_i^\mu \in D_\mu \wedge \max(p_m(y|\theta_t^\mu(X_i^\mu))) > \mathcal{T}_{t+1}(\hat{y}_i^\mu)\}_{i=1}^{N_\mu}$$

threshold of class c

$$\tau_{t,c} = \widetilde{p}_t(c) + \tau - std(\widetilde{p}_t)$$

Theorem 3.1.

$$\tau + \widetilde{p}_t(c) - \sqrt{\frac{1}{C}} \leq \mathcal{T}_t(c) \leq \tau + \widetilde{p}_t(c).$$

Since $\tau >> \sqrt{\frac{1}{C}}$, $\mathcal{T}_t(c)$ will have a high lower bound

modified $\widetilde{p}_t(c)$

$$\widetilde{p}_t(c) = \frac{\sigma_t(c)}{\sum_{i=1}^{C}\sigma_t(i)} \times \frac{C}{10}.$$

Method

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

• Discovery of Unlabeled Data from Tail Classes

Forwarm up stage in labeled clients, it is similar to long-tailed classification, so the problems in long-tailed classification will also exist in our pseudo labeling process : models tend to classify tail (rare) classes as head (common) classes

mask function

$$\mathcal{M}_i(p) = \begin{cases} p_i & i \neq \arg\max p \\ 0 & i = \arg\max p \end{cases}$$

analyze the second largest confidence score

$$\hat{y}_i^{u\prime} = \arg\max \mathcal{M}(p_m(y|\theta_t^\mu(X_i^\mu)))$$

misclassfied data

$$D_\mu^{tail} = \{(X_i^\mu, \hat{y}_i^{\mu\prime}) | X_i^\mu \in D_\mu$$
$$\wedge \max(p_m(y|\theta_t^\mu(X_i^\mu))) \leq \mathcal{T}_t(\hat{y}_i^\mu) \wedge \widetilde{p}_t(\hat{y}_i^{\mu\prime}) < \frac{\beta}{C}\}$$

$$D_\mu^{train} = D_\mu^{tail} \cup \widetilde{D}_\mu$$

$$\mathcal{L}_\mu = \frac{1}{|D_\mu^{train}|} \sum_{(X^\mu, \hat{y}^\mu) \in D_\mu^{train}} H(\hat{y}^\mu, p_m(y|\theta_t^\mu(X^\mu))).$$

$$\sigma_t^\mu(c) = \sum_{(X^\mu, \hat{y}^\mu) \in D_\mu^{train}} \mathbf{1}(\hat{y}^\mu = c).$$

Method

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

- **Aggregation of local models**

$$w_t^i = \begin{cases} \frac{|D_i|}{|D_t^{train}|} & \text{if } i \in \{1, \cdots, m\} \\ \frac{|D_{t,i}^{train}|}{|D_t^{train}|} & \text{if } i \in \{m+1, \cdots, m+n\} \end{cases}$$

$$|D_t^{train}| = \sum_{\ell=1}^{m} |D_\ell| + \sum_{\mu=m+1}^{m+n} |D_{t,\mu}^{train}|$$

$$\theta_{t+1}^G = \begin{cases} \sum_{i=1}^{m+n} w_t^i \theta_t^i & t\%s \neq 0 \\ \alpha_2 \theta_{t+1-s}^G + (1-\alpha_2) \sum_{i=1}^{m+n} w_t^i \theta_t^i & t\%s = 0, \end{cases}$$

Experiments

Table 1. Results on SVHN, CIFAR-10/100, Fashion MNIST and ISIC 2018 datasets under heterogeneous data partition with ResNet18. FedAVG$^+$ means FedAvg [19] trained with all one labeled clients using our residual weight connection. Fed-consist$^+$ means Fed-Consist [31] using our proposed fixed pseudo labeling without enlarging the weight of labeled client.

Labeling Strategy	Method	Client Num.		Dataset				
		labeled	unlabeled	SVHN	CIFAR10	CIFAR100	Fashion-MNIST	ISIC 2018
Fully supervised	FedAvg [19](upper-bound)	10	0	91.83	80.89	51.38	90.14	81.32
	FedAvg [19](lower-bound)	1	0	67.71	54.66	20.49	74.87	65.13
	FedAvg$^+$ [19]	1	0	76.98	58.21	24.84	78.26	66.69
Semi supervised	FedIRM [18]	1	9	69.22	52.84	20.20	76.83	64.85
	Fed-Consist [31]	1	9	70.56	54.23	21.81	76.57	65.20
	Fed-Consist$^+$ [31]	1	9	86.57	56.35	23.25	78.35	65.50
	RSCFed [14]	1	9	76.74	57.07	28.46	78.40	67.21
	CBAFed(ours)	1	9	**88.07**	**67.08**	**30.18**	**85.49**	**68.29**

local training epoch： 11 (labeled client) / 1 (unlabeled client)

Experiments

Table 2. Comparison of our method against RSCFed [14], Fed-Consist [31] and FedAVG [19] in SVHN dataset on ViT [5] as the backbone, with one labeled and nine unlabeled clients.

Method	Client Num.		Accuracy
	labeled	unlabeled	
FedAVG [19](upper bound)	10	0	96.81
FedAVG [19](lower bound)	1	0	81.68
FedAVG$^+$ [19]	1	0	88.93
FedIRM [18]	1	9	79.44
Fed-Consist [31]	1	9	85.91
Fed-Consist$^+$ [31]	1	9	93.21
RSCFed [14]	1	9	89.43
CBAFed(ours)	1	9	**95.09**

Table 3. Comparison of our method against RSCFed [14], Fed-Consist [31], FedIRM [18] and FedAVG [19] with the number of labeled and unlabeled client set to 2 and 8.

Method	Client Num.		Accuracy
	labeled	unlabeled	
FedAVG [19](upper bound)	10	0	80.89
FedAVG [19](lower bound)	2	0	61.85
FedAVG$^+$ [19]	2	0	66.55
FedIRM [18]	2	8	62.62
Fed-Consist [31]	2	8	61.67
Fed-Consist$^+$ [31]	2	8	68.04
RSCFed [14]	2	8	64.25
CBAFed(ours)	2	8	**72.01**

Experiments

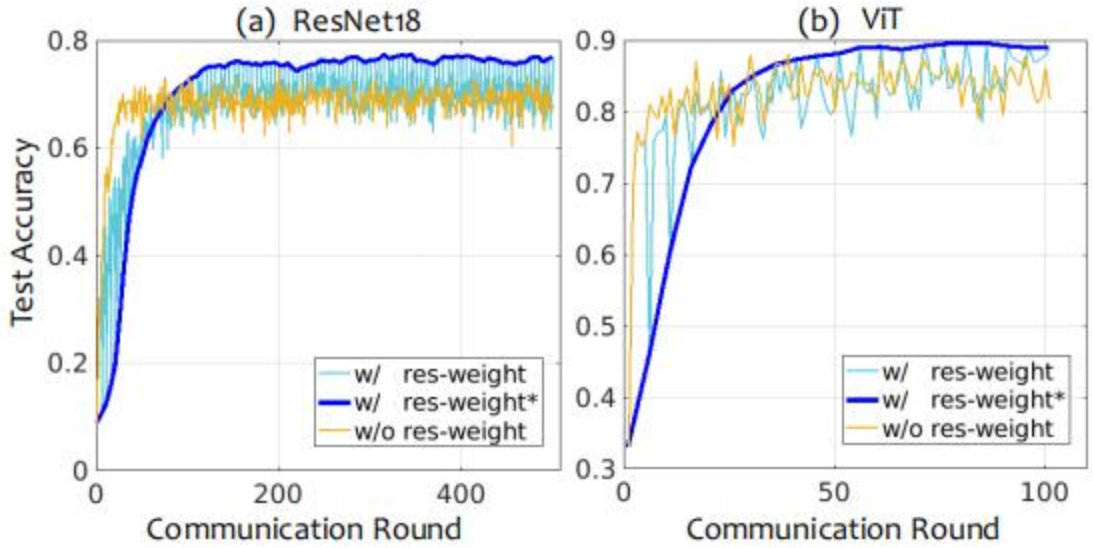


Figure 2. Test accuracy curves in local training of SVHN dataset w/ and w/o residual weight connection. ResNet18 (a) and ViT (b) are adopted as the backbones. W/ res-weight* indicates we only show test accuracy on epochs (communication rounds, since local training epoch for labeled client is 1) with skip weight connection. Best viewed electronically.

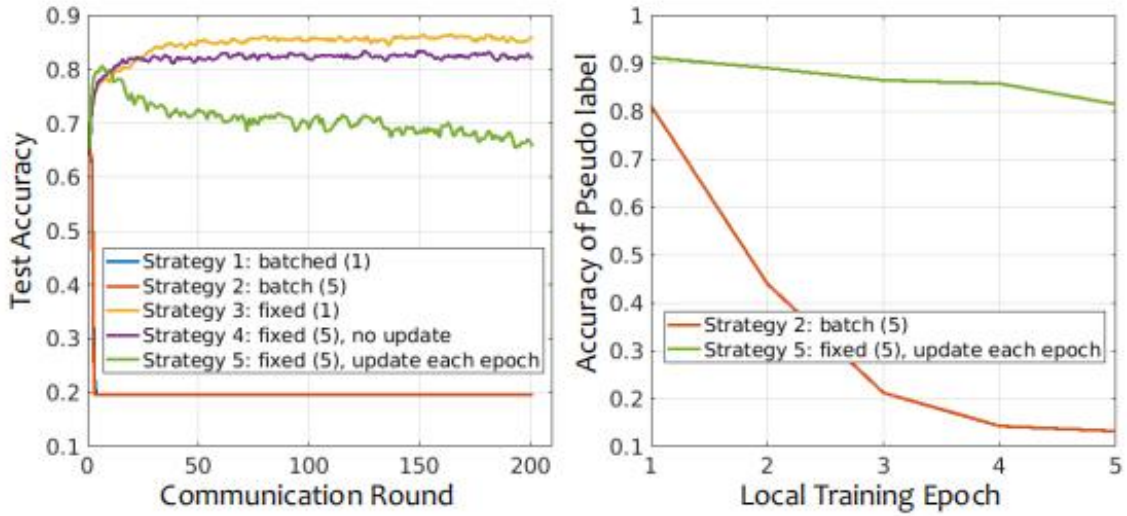


Figure 3. Left: Test accuracy of all 5 strategies after every communication round. Note that the test accuracy of communication round 0 is the test accuracy of model trained on labeled client. Right: Accuracy of pseudo labels in local training epoch of one randomly selected unlabeled client. Best viewed electronically.

Experiments

Table 4. Ablation Study of CBAFed in CIFAR-10/100 and Fashion MNIST Datasets. Fixed PL: fixed pseudo labeling, CBA: class balanced adaptive pseudo labeling, DD: tail class data discovery.

Dataset	Fixed PL	CBA	DD	Res-Weight	Accuracy
CIFAR-10	✓				59.16
	✓	✓			64.29
	✓	✓	✓		65.15
	✓	✓	✓	✓	67.08
CIFAR-100	✓				27.64
	✓	✓			29.41
	✓	✓	✓		29.86
	✓	✓	✓	✓	30.18
Fashion-MNIST	✓				79.99
	✓	✓			80.87
	✓	✓	✓		84.37
	✓	✓	✓	✓	85.49

Experiments

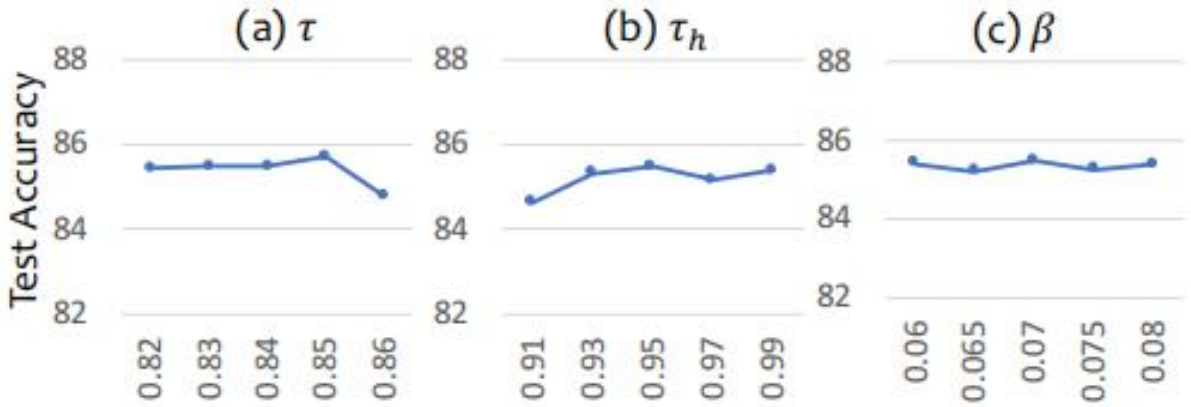


Figure 4. Performance changes on Fashion-MNIST by varying (a) threshold base τ and (b) upper bound threshold τ_h, and (c) parameter for selecting tail class data β.

Thanks