How to Prevent the Poor Performance Clients for Personalized Federated Learning?

Zhe Qu[1], Xingyu Li[2], Xiao Han[3], Rui Duan[3], Chengchao Shen[1], Lixing Chen[4]
[1]Central South University, [2]Mississippi State University,
[3]University of South Florida, [4]Shanghai Jiaotong University

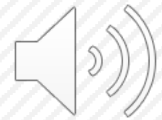
01 Background

➢User Modeling

- ➢ An important basis to capture useful potential characteristics with the reliance on personal data
- ➢ User modeling has been applied to multiple typical, such as modeling capabilities to preferences from users
- ➢ User modeling processes usually centralized training with data aggregated, which causes privacy leakage

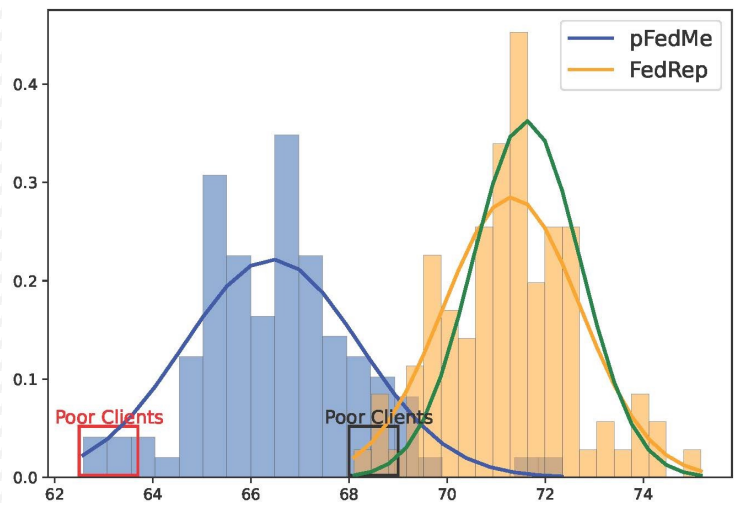
➢Federated Leaning/Personalized Federated Learning

- ➢ Federated Leaning (FL) refers to building and aggregating user models while leaving private data isolated so that preserves the data privacy
- ➢ Due to limited local data samples, Personalized Federated Learning (pFL) can obtain customize local models to fit the local dataset



02 Problems and Design Intuition

- Existing pFL studies use averaged learning accuracy to evaluate their proposed algorithms.
- Existing pFL studies cannot efficiently protect the clients with poor performance.
- Existing pFL studies can be divided by two schemes: with or without global model.



These poor clients may be generated by the biased learned universal information.
Hence, we aim to design a strategy to moderate this biased phenomenon.

Contributions:

- Present the reason why we need to focus on the poor performance clients.
- Design a personalize locally generalize universally strategy to moderate the biased universal information.
- Propose three PLGU-based algorithms on two pFL schemes.

03

Personalize Locally Generalize Universally (PLGU)

Personalization score $\xi_{i,l}^t$ of the scaling matrix $\boldsymbol{\Lambda}_i$ is

$$\xi_{i,l}^t = \frac{\|\tilde{\boldsymbol{\theta}}_{i,l}^t - \widetilde{\boldsymbol{w}}_l^t\|}{\dim(\widetilde{\boldsymbol{w}}_l^t)}$$

$\tilde{\boldsymbol{\theta}}_{i,l}^t$ and $\widetilde{\boldsymbol{w}}_l^t$ are the model parameters at the l-th layer of the personalized model $\tilde{\boldsymbol{\theta}}_{i,l}^t$ and the global model $\widetilde{\boldsymbol{w}}_l^t$.

$\dim(\bullet)$ denotes the number of parameters on layer l, which can normalize the values as $\sum_{l=1}^{L} \xi_{i,l}^t = 1$

03 Personalize Locally Generalize Universally (PLGU)

Layer-Wised Sharpness Aware Minimization (LWSAM):

Let $\boldsymbol{\Lambda}_i$ denote a diagonal $L \times L$ matrix, $\boldsymbol{\Lambda}_i = \mathrm{diag}(\xi_{i,1}, \ldots, \xi_{i,L})$, where $\xi_{i,l}$ is the layer personalization score. We apply the adopted scaling method to the inner maximization of SAM

$$F_{\mathcal{D}_i}(\widetilde{\boldsymbol{\theta}}_i) = \max_{\|\boldsymbol{\Lambda}_i \boldsymbol{\epsilon}_i\| \leq \rho} F_{\mathcal{D}_i}(\boldsymbol{\theta}_i + \boldsymbol{\Lambda}_i \boldsymbol{\epsilon}_i),$$

where $\boldsymbol{\theta}_i$ is personalized model, ρ is the radius, $\widetilde{\boldsymbol{\theta}}_i = \boldsymbol{\theta}_i + \boldsymbol{\epsilon}_i$. The approximate inner solution of LWSAM is

$$\widetilde{\boldsymbol{\epsilon}}_i = \rho sign(\nabla_{\boldsymbol{\theta}_i} F_{\mathcal{D}_i}(\boldsymbol{\theta}_i)) \boldsymbol{\Lambda}_i \frac{\nabla_{\boldsymbol{\theta}_i} F_{\mathcal{D}_i}(\boldsymbol{\theta}_i)}{\|\nabla_{\boldsymbol{\theta}_i} F_{\mathcal{D}_i}(\boldsymbol{\theta}_i)\|}$$

It provides the layer-wise calculate of $\widetilde{\boldsymbol{\epsilon}}_i$ to scale up the batch size on client i.

Personalize Locally Generalize Globally

Algorithm 1 PLGU($\tilde{\boldsymbol{\theta}}_i^{t,0}, \tilde{\boldsymbol{w}}_i^{t,0}, \boldsymbol{\Lambda}_i^t, K, \eta$).

Input: personalized model $\tilde{\boldsymbol{\theta}}_i^{t,0}$, global model $\tilde{\boldsymbol{w}}_i^{t,0}$, scaling matrix $\boldsymbol{\Lambda}_i^t$, number of local epochs K, learning rate η;
for $k = 0, \ldots, K-1$ **do**
  Sample mini-batch $\mathcal{B}_i$ on client i;
  Calculate unbiased gradient $\boldsymbol{g}_i^{t,k} = \nabla_{\tilde{\theta}_i^{t,k}} F_{\mathcal{B}_i}(\tilde{\boldsymbol{\theta}}_i^{t,k})$;
  Update personalized model $\tilde{\boldsymbol{\theta}}_i^{t,k+1} = \tilde{\boldsymbol{\theta}}_i^{t,k} - \eta \boldsymbol{g}_i^{t,k}$;
  Calculate perturbation $\tilde{\boldsymbol{\epsilon}}_i^{t,k}$ by (5);
  Calculate unbiased gradient approximation for LWSAM $\tilde{\boldsymbol{g}}_i^{t,k} = \nabla_{\tilde{\boldsymbol{w}}_i^{t,k} + \tilde{\boldsymbol{\epsilon}}_i^{t,k}} F_{\mathcal{B}_i}(\tilde{\boldsymbol{\theta}}_i^{t,k} + \tilde{\boldsymbol{\epsilon}}_i^{t,k})$;
  Update global model $\tilde{\boldsymbol{w}}_i^{t,k+1} = \tilde{\boldsymbol{w}}_i^{t,k} - \eta \tilde{\boldsymbol{g}}_i^{t,k}$;
end for

04 PLGU-Layer Freezing (PLGU-LF)

The objective of PLGU-LF is:

$$\min_{\boldsymbol{w},\{\boldsymbol{\theta}_i\}_{i=1}^N} \max_{\{\boldsymbol{\Lambda}_i \boldsymbol{\epsilon}_i \le \rho\}_{i=1}^N} \frac{1}{N} \sum_{i=1}^{N} F_i(\widetilde{\boldsymbol{w}}; \widetilde{\boldsymbol{\theta}}_i)$$

where $\widetilde{\boldsymbol{w}} = \boldsymbol{w} + \widetilde{\boldsymbol{\epsilon}}_i$, $\widetilde{\boldsymbol{\theta}}_i = \boldsymbol{\theta}_i + \widehat{\boldsymbol{\epsilon}}_i$

The generalization bound of PLGU-LF is:

$$O\Bigg(\beta\Bigg(\frac{B\sqrt{L}\prod_{l\in\mathcal{C}} M_l}{\gamma\sqrt{m}} + \frac{(B+\rho)d\sqrt{L\log L}\prod_{l\in\mathcal{C}} M_l}{\gamma\sqrt{m}} + \frac{B\sqrt{L}\prod_{l\in\mathcal{C}} M_l}{\gamma\sqrt{m}} + \frac{dD(L-)\sqrt{\log D}\prod_{l\in\mathcal{C}^{Per}} BM_l\sqrt{\log(\mathrm{L}-\mathrm{D})}\prod_{l\in\mathcal{C}^{Uni}}(B+\rho)M_l}{\gamma N\sqrt{m}}\Bigg) + \sqrt{B\log 1/\gamma}\Bigg)$$

Algorithm 2 Scheme I: PLGU-LF algorithm.

Input: communication upper bound T, client set $\mathcal{N}$, number of local epochs K, learning rate η;
Output: personalized model $\tilde{\boldsymbol{w}}_i^T$ and global model $\tilde{\boldsymbol{w}}^T$;
for $t = 0, \ldots, T-1$ **do**
  Sample a set of clients $\mathcal{C}^t \subseteq \mathcal{N}$ with the size of C;
  for each client $i \in \mathcal{C}^t$ in parallel **do**
    Calculate $\boldsymbol{\Lambda}_i^t$ by (6);
    Select D layers with largest ξ_i^t values to be the set as $\mathcal{L}_{i,\text{Per}}^t$ and other layers are set as $\mathcal{L}_{i,\text{Uni}}^t$;
    $\tilde{\boldsymbol{\theta}}_{i,\mathcal{L}^{\text{Uni}}}^{t,0} = \tilde{\boldsymbol{w}}_{\mathcal{L}^{\text{Uni}}}^t$ and $\tilde{\boldsymbol{\theta}}_{i,\mathcal{L}^{\text{Per}}}^{t,0} = \tilde{\boldsymbol{\theta}}_{i,\mathcal{L}^{\text{Per}}}^t$
    PLGU$(\tilde{\boldsymbol{\theta}}_i^{t,0}, \tilde{\boldsymbol{w}}_i^{t,0}, \boldsymbol{\Lambda}_i^t, K, \eta)$
    $\Delta_i^t = \tilde{\boldsymbol{w}}_i^{t,K} - \tilde{\boldsymbol{w}}_i^{t,0}$;
    $\tilde{\boldsymbol{w}}^{t+1} = \tilde{\boldsymbol{w}}^{t+1} + \frac{1}{C}\sum_{i\in\mathcal{C}^t} \Delta_i^t$;
  end for
end for

PLGU-Generalized Representation (PLGU-GRep)

The objective of PLGU-GRep is:

$$\min_{\boldsymbol{\phi}\in\Phi} \max_{\{\boldsymbol{\Lambda}_i \boldsymbol{\epsilon}_i \le \rho\}_{i=1}^{N}} \frac{1}{N}\sum_{i=1}^{N} \max_{\boldsymbol{h}_i \in \mathcal{H}} F_i(\boldsymbol{h}_i, \boldsymbol{\phi})$$

The generalization bound of PLGU-GRep is:

$$O\left(\beta\left(\frac{\beta_{\boldsymbol{\phi}}(B+\rho)d\sqrt{(L-1)\log(L-1)}\prod_{l\in\mathcal{L}_{\boldsymbol{\phi}}}M_l}{\gamma\sqrt{m}}+\frac{\beta_{\boldsymbol{h}}B\sqrt{L}M_h}{\gamma N\sqrt{m}}\right.\right.$$
$$\left.\left.+\frac{\beta_{\boldsymbol{\phi}}(B+\rho)d\sqrt{(L-1)\log(L-1)}\prod_{l\in\mathcal{L}_{\boldsymbol{\phi}}}M_l}{\gamma N\sqrt{m}}\right)+\sqrt{m\log 1/\gamma}\right)$$

Algorithm 3 Scheme II: PLGU-GRep algorithm.

Input: communication upper bound T, client set $\mathcal{N}$, number of header local epochs K, learning rate η;
Output: personalized model $\tilde{\boldsymbol{\theta}}_i^T$;
for $t = 0, \ldots, T-1$ **do**
  Sample a set of clients $\mathcal{C}^t \subseteq \mathcal{N}$ with the size of C;
  Download global representation $\tilde{\boldsymbol{\phi}}^t$ to client $i \in \mathcal{C}^t$;
  for each client $i \in \mathcal{C}^t$ in parallel **do**
    for $k = 0, \ldots, K-1$ **do**
      Sample mini-batch $\mathcal{B}_i \subset \mathcal{D}_i$;
      Update the header $\boldsymbol{h}_i^{t,k}$ by (9);
    end for
    Sample mini-batch $\mathcal{B}_i \subset \mathcal{D}_i$;
    Update the generalized representation $\tilde{\boldsymbol{\phi}}_i^t$ by (10)-(13);
  end for
  Server updates the new representation $\tilde{\boldsymbol{\phi}}^{t+1} = \frac{1}{C}\sum_{i\in\mathcal{C}^t}\tilde{\boldsymbol{\phi}}_i^t$;
end for

PLGU-Generalized Hypernetwork (PLGU-GHN)

The objective of PLGU-GHN is:

$$\min_{\boldsymbol{\phi}\in\Phi,\{\boldsymbol{v}_i\}_{i=1}^N} \max_{\{\boldsymbol{\epsilon}\le\rho\}_{i=1}^N} \frac{1}{N}\sum_{i=1}^{N} F_i(h(\boldsymbol{v}_i, \boldsymbol{\phi}+\boldsymbol{\epsilon}))$$

The generalization bound of PLGU-GRep is:

$$O\left(\beta\left(\frac{\beta_{\boldsymbol{v}} B\sqrt{L}\prod_{l\in\mathcal{C}} M_l}{\gamma\sqrt{m}} + \frac{\beta_{\boldsymbol{h}} R\sqrt{L_h}\prod_{l\in\mathcal{C}_h} M_l^h}{\gamma N\sqrt{m}} + \frac{(R+\rho)d_h\sqrt{L_h\log L_h}\prod_{l\in\mathcal{C}_h} M_l^h}{\gamma\sqrt{N}}\right) + 2\sqrt{\frac{\log 1/\gamma}{N}}\right)$$

Algorithm 4 Scheme II: PLGU-GHN algorithm.

Input: communication upper bound T, client set $\mathcal{N}$, number of local epochs K, learning rate η;
Output: personalized model $\boldsymbol{w}_i^T$;
for $t = 0, \ldots, T-1$ **do**
  Sample a set of clients $\mathcal{C}^t \subseteq \mathcal{N}$;
  for each client $i \in \mathcal{C}^t$ in parallel **do**
    set $\boldsymbol{w}_i^t = h(\boldsymbol{v}_i^t; \boldsymbol{\phi}^t)$ and $\hat{\boldsymbol{w}}_i^t = \boldsymbol{w}_i^t$;
    for $k = 0, \ldots, K-1$ **do**
      sample mini-batch $\mathcal{B}_i \subset \mathcal{D}_i$;
      $\boldsymbol{w}_i^{t,k+1} = \boldsymbol{w}_i^{t,k} - \eta\nabla_{\boldsymbol{w}_i^{t,k}} F_{\mathcal{B}_i}(\boldsymbol{w}_i^{t,k})$;
    end for
  end for
  $\Delta\boldsymbol{w}_i^t = \boldsymbol{w}_i^{t,K} - \boldsymbol{w}_i^{t,0}$;
  Calculate $\boldsymbol{g}^t$, $\tilde{\boldsymbol{\phi}}^t$, and $\boldsymbol{v}_i^t$ by (24) and (25);
end for

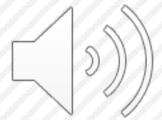
05 Experiments

Table 1. Testing accuracy by the global model (averaged, top 5%, and lowest 5% accuracy) under three datasets.

Datasets	C	FedAvg	FedSAM	Ditto	pFedMe	PLGU-LF
CIFAR10	10	64.79	67.57	64.15	64.36	69.36
		60.69 68.45	64.76 69.93	62.08 69.45	61.13 70.46	67.34 71.69
	100	67.15	69.49	68.24	68.31	71.48
		65.99 71.80	67.04 71.12	61.64 71.08	65.25 71.64	68.92 73.57
CIFAR100	10	55.86	57.19	56.35	55.17	59.74
		51.21 60.75	54.82 59.55	52.73 59.28	50.41 58.35	56.85 61.49
	100	58.00	59.39	58.93	57.24	61.07
		55.16 60.67	56.73 61.20	56.64 60.79	52.96 60.53	58.25 62.25
TmgNet	20	37.78	38.42	38.09	36.43	40.61
		32.26 43.73	34.61 42.02	34.75 42.84	31.79 42.90	35.41 43.56

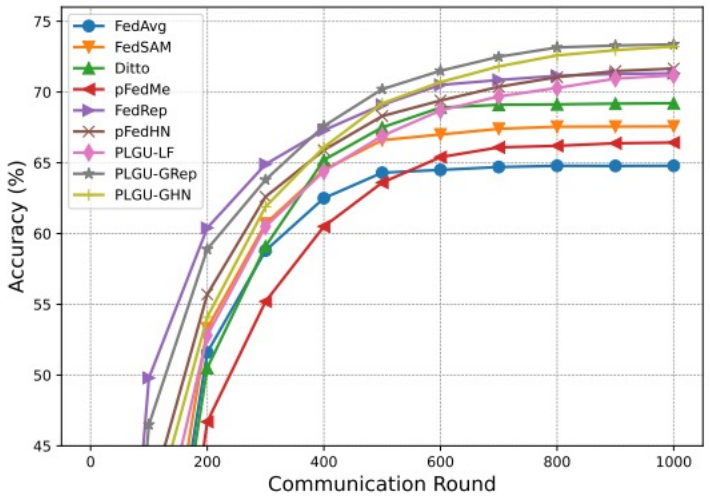
Table 2. Testing accuracy by the personalized model (averaged, top 5%, and lowest 5% accuracy) under three datasets.

Datasets	C	Ditto	pFedMe	FedRep	pFedHN	PLGU-LF	PLGU-GRep	PLGU-GHN
CIFAR10	10	69.21	66.43	71.32	71.66	71.18	72.87	72.64
		65.33 72.86	62.97 71.10	68.15 74.04	68.06 73.95	68.22 73.27	69.98 74.91	69.70 74.63
	100	71.23	69.19	75.30	74.95	74.23	76.94	76.17
		69.93 74.46	66.84 71.58	73.11 77.74	73.23 77.58	72.66 75.87	74.23 77.61	73.79 77.82
CIFAR100	10	59.17	56.58	62.92	63.25	62.33	64.61	65.37
		57.81 63.06	52.92 60.14	59.70 65.95	59.86 65.79	60.08 65.24	62.58 66.62	63.02 66.94
	100	62.52	58.57	65.08	65.54	64.67	66.79	66.30
		59.48 64.81	55.79 61.73	62.60 67.96	63.30 68.09	63.72 67.75	64.81 68.59	64.27 68.02
TmgNet	20	39.41	37.22	41.68	41.96	41.39	42.84	42.45
		35.88 43.49	32.76 42.91	37.53 45.07	37.94 45.19	37.46 44.57	40.29 45.18	39.92 44.89

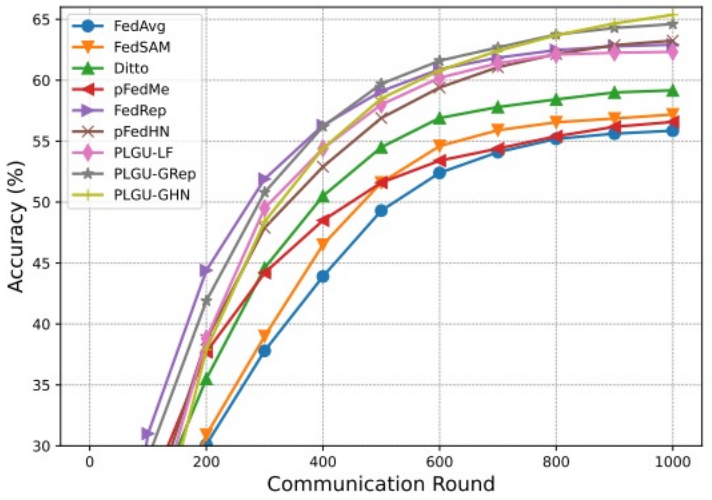


05

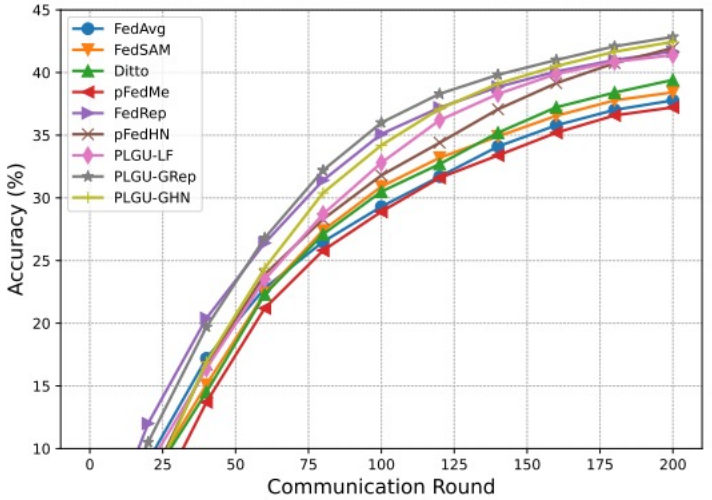
Experiments



(a) CIFAR10.



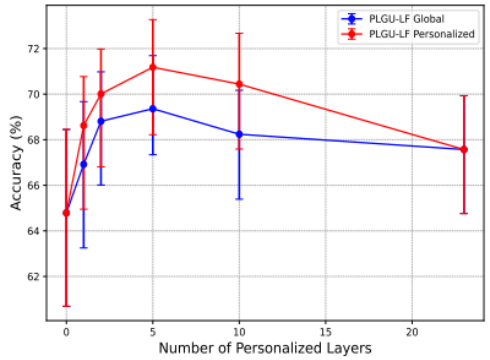
(b) CIFAR100.



(c) TmgNet.

05

Experiments



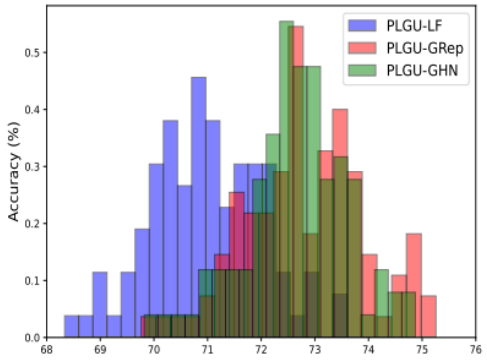


Table 3. Impact of number of local epochs K with $C = 10$.

K	1		5		10	
PLGU-LF (G)	65.31		69.36		68.42	
	62.08	68.20	67.24	71.13	66.23	71.39
PLGU-LF (P)	67.43		71.18		70.40	
	64.25	70.01	68.22	73.17	65.76	73.54
PLGU-GRep	70.82		72.87		71.65	
	67.25	73.28	69.98	74.91	68.04	74.11
PLGU-GHN	69.93		72.64		72.91	
	66.89	73.15	69.70	74.63	72.91	75.48

Table 4. Impact of perturbation ρ with $C = 10$.

ρ	0.05		0.1		0.5	
PLGU-LF (G)	69.36		68.62		66.73	
	67.24	71.13	65.71	71.88	63.95	68.31
PLGU-LF (P)	71.18		70.96		69.05	
	68.22	73.17	69.15	74.68	66.89	73.50
PLGU-GRep	72.87		72.01		70.80	
	69.98	74.91	71.49	75.93	66.73	73.65
PLGU-GHN	72.64		70.49		67.30	
	69.70	74.63	68.18	74.73	64.84	74.69

Experiments

Thank you!