DM-PFL: Hitchhiking Generic Federated Learning for Efficient Shift-Robust Personalization

Wenhao Zhang[1], Zimu Zhou[2], Yansheng Wang[1], Yongxin Tong[1]

[1]State Key Laboratory of Software Development Environment, Beijing Advanced Innovation Center for Future Blockchain and Privacy Computing, Beihang University, Beijing, China

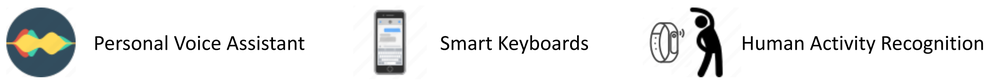
[2]City University of Hong Kong, Hong Kong SAR, China

[1]{garyzhang, arthur_wang, yxtong}@buaa.edu.cn

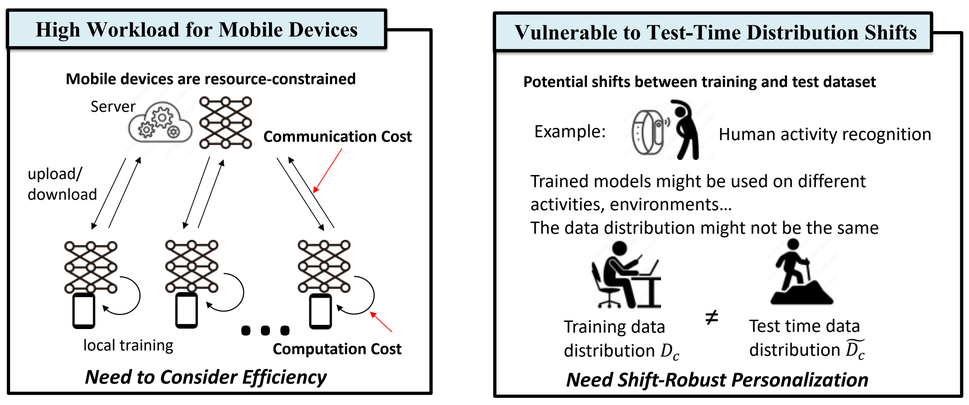
[2]zimuzhou@cityu.edu.hk

Introduction

- Unlike the generic federated learning that learns a global model for all clients, personalized federated learning trains client-specific models collaboratively, which holds potential for various mobile applications with heterogeneous data.



- Despite extensive personalized FL proposals two drawbacks impede their practical usage for mobile applications.



- In response, we explore efficient shift-robust personalization for FL, which aims to train personalized models that are robust to test-time distribution shifts with low training overhead.

DM-PFL Overview

Review on Generic and Personalized FL

Generic FL:
- Train a global model with better generalization ability that performs well under a wider range of data distributions.

$$\min_{\theta_g} l(\theta_g) = \sum_{c=1}^{C} \frac{|D_c|}{|D|} \mathbb{E}[\mathcal{L}(\theta_g; D_c)]$$

θ_g : the global model

Optimizing performance on the aggregation of large amounts of data across clients.

- **Good local performance**
- **Vulnerable to distribution shifts.**

Personalized FL:
- Train client-specific personalized models for better performance on client's local data distributions.

$$\min_{\theta_1,\theta_2,\ldots,\theta_C} l(\theta_1, \ldots, \theta_C) = \sum_{c=1}^{C} \frac{|D_c|}{|D|} \mathbb{E}[\mathcal{L}(\theta_c; D_c)]$$

θ_c: each client's local model

Optimize performance on the clients' local data.

- **Sub-optimal local performance**
- **Robust to distribution shifts.**

Design Rationales:

1. Hitchhike the global model (θ_g) to make our personalized models (θ_c) more shift-robust.
2. Use mask-based training to enforce model sparsity and enable weight-level parameter sharing.

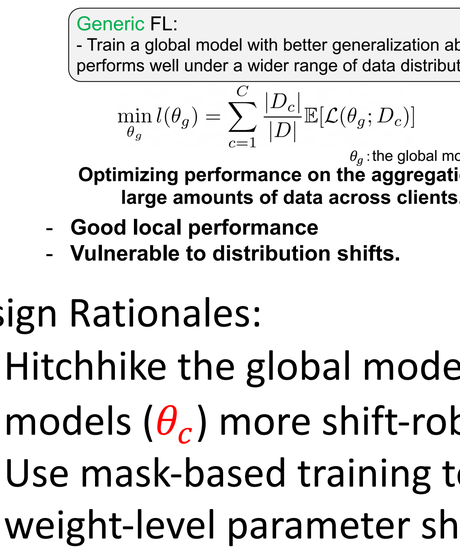
Our objectives and constraints:

$$\min_{\theta_1,\theta_2,\ldots,\theta_C} l(\theta_1, \ldots, \theta_C) = \sum_{c=1}^{C} \frac{|D_c|}{|D|} \mathbb{E}[\mathcal{L}(\theta_c; \tilde{D}_c)]$$
$$s.t. \quad \|\theta_c\|_0 \leq |\theta_c| * (1 - \mathbb{S}), \forall c \in \{1, 2, \ldots, C\}$$

Objective — Take test-time distribution shift $\widetilde{D_c}$ into consideration

Constraint — Efforcing model sparsity to improve efficiency

Dual Masking Mechanism



Components

We use global mask m_g and a set of personalized masks $\{m_c\}_{c=1}^{C}$ with weights w_g and w_c to construct dual models.

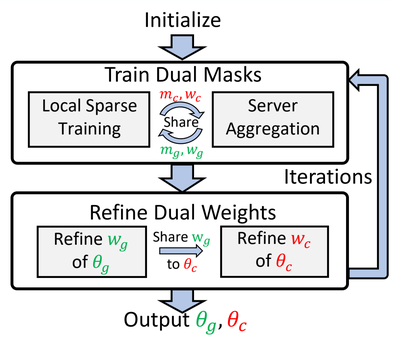
$$\boldsymbol{\theta_g = w_g \odot m_g}$$
$$\boldsymbol{\theta_c = w_g \odot (m_g \cap m_c) + w_c \odot (m_c - m_g)}$$

- Each client inherits a different portion of the global weights w_g according to its overlap in mask positions with the global mask, *i.e.*, $m_g \cap m_c$.
- For the remaining positions $m_c - m_g$ in its personalized mask, each client inherits weights from its own w_c.

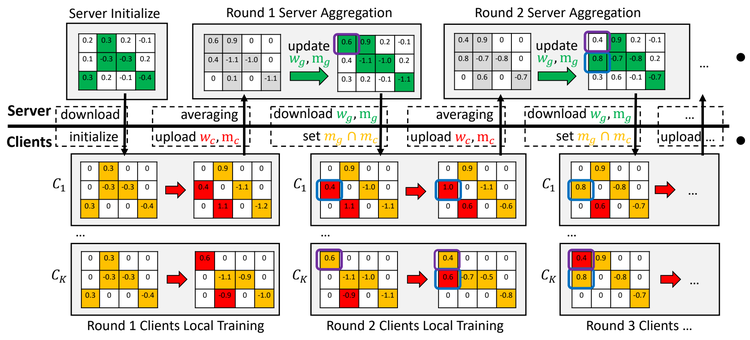
DM-PFL Algorithm

Two-staged sparse training algorithm

- In the first train dual masks stage, we train weights and masks of the dual models interactively.
- In the refine dual weights stage, we first refine w_g then refine w_c under fixed masks.



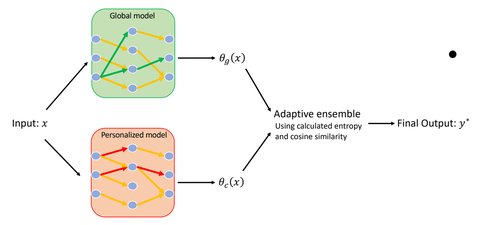
Example of train dual masks stage



- Dual masks and weights are trained interactively.
- Shared weights can transform into personalized weights(purple box), and vice versa(blue box).

- With our dual masking design, training the global weights can be viewed as train the shared weights of personalized model of each client, so the workload of our training scheme is much less than training dual models independently.

Adaptive inference



- We may further utilize θ_g to improve shift-robustness, which adaptively ensembles θ_g and θ_c based on the estimated degree of test-time shift.

Experimental Evaluation

Performance with and without shifts

Different Non-IID settings

Non-IID	Dir(0.1)		Pathological	
Methods	w/o	w/ 100%	w/o	w/ 100%
FedAvg+FT	89.00±11.89	23.85±25.32	89.76±6.78	20.47±31.26
LG-FedAvg	85.24±13.90	17.38±24.76	84.41±7.34	15.54±31.69
FedPer	88.82±12.63	22.76±30.10	89.17±6.84	16.91±34.98
FedRep	89.19±10.48	21.60±27.40	89.69±6.15	17.24±34.97
Ditto	88.20±9.12	39.79±26.91	89.22±5.53	22.46±28.03
APFL	89.18±9.76	22.15±29.28	89.41±6.87	21.48±34.69
DM-PFL	**89.29±11.65**	45.24±23.54	**90.06±5.59**	48.95±24.20
DM-PFL+	88.16±12.46	**47.33±16.53**	88.59±7.65	**54.13±23.44**

Dirichlet distribution-based CIFAR10 — **Pathological CIFAR10**

Dataset	HAR		FEMNIST	
Methods	w/o	w/ 100%	w/o	w/ 100%
FedAvg+FT	97.38±5.53	92.84±6.97	98.85±5.23	96.86±9.76
LG-FedAvg	98.16±9.12	80.12±11.23	93.15±8.86	62.44±19.70
FedPer	98.07±3.54	94.52±5.50	98.75±6.78	97.44±5.61
FedRep	97.91±3.11	90.14±8.22	96.49±6.42	84.16±13.97
Ditto	97.52±4.03	94.53±6.17	98.71±4.48	96.72±7.31
APFL	95.51±5.62	87.47±10.19	98.49±3.76	96.55±8.05
DM-PFL	**98.41±2.72**	95.37±4.05	**99.04±3.80**	97.86±5.28
DM-PFL+	97.93±3.46	**95.58±4.76**	98.62±4.63	**98.10±5.01**

Realistic HAR — **Realistic FEMNIST**

Communication and computation costs

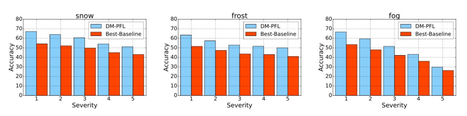
Methods	Comm (MB)		FLOP (1e12)	
Local	0	0%	6.12	100%
FedAvg	90.75	100%	6.12	100%
FedProx	90.75	100%	6.12	100%
FedDST	46.44	51.2%	3.90	63.7%
FedAvg+FT	90.75	100%	6.32	103%
LG-FedAvg	60.71	66.9%	6.12	100%
FedPer	90.53	99.7%	6.12	100%
FedRep	90.53	99.7%	8.28	131%
Ditto	90.75	100%	12.27	201%
APFL	90.75	100%	12.26	200%
DM-PFL	36.58	40.3%	3.91	63.9%

Different Non-IID settings

Type	Methods	w/o shift	w/ 20% shift	w/ 40% shift	w/ 60% shift	w/ 80% shift	w/ 100% shift	Averaged
Local	Local	61.86±9.53	54.11±11.66	46.23±13.78	38.38±15.84	30.70±17.97	22.93±20.07	42.37±14.81
GFL	FedAvg	60.16±8.21	59.84±8.84	59.53±9.49	59.20±10.14	58.88±10.78	58.56±11.42	59.36±9.81
	FedProx	58.61±10.60	58.52±10.55	58.43±10.5	58.33±10.46	58.24±10.41	58.15±10.36	58.38±10.48
	FedDST	60.13±6.76	59.25±8.70	58.89±8.84	58.52±8.97	58.15±9.11	57.79±9.24	58.79±8.60
PFL	FedAvg+FT	79.33±8.68	72.12±10.72	65.34±12.68	58.41±14.67	51.50±16.69	44.60±18.68	61.88±13.69
	LG-FedAvg	69.50±13.29	60.53±14.52	51.09±15.76	41.83±16.96	32.80±18.20	23.62±19.42	46.56±16.36
	FedPer	76.57±12.27	68.08±14.45	59.69±16.53	51.02±18.67	42.54±20.74	34.13±22.87	55.34±17.59
	FedRep	76.85±11.31	67.40±13.00	58.49±14.80	49.28±16.51	40.24±18.24	31.20±19.95	53.91±15.63
	Ditto	79.70±8.59	73.17±10.24	67.56±11.68	61.40±13.22	55.11±14.74	49.04±16.25	64.33±12.45
	APFL	78.41±9.92	69.97±11.81	62.03±13.78	53.50±15.75	45.42±17.69	37.13±19.63	57.74±14.76
Ours	DM-PFL	**79.93±8.61**	**75.36±9.55**	**70.75±10.53**	66.09±11.46	61.38±12.39	56.79±13.34	68.38±10.98
	DM-PFL+	77.22±9.31	73.81±8.80	70.41±8.30	**66.97±7.79**	**63.61±7.27**	**59.61±8.57**	**68.61±8.34**

CIFAR10

Different types of shifts



(a) snow (b) frost (c) fog

CIFAR10-C

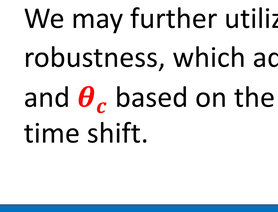
Ablation studies

(a) Impact of sparsity on communication cost (b) Impact of sparsity on computation cost (c) Impact of sparsity on accuracy w/o shift

(d) Impact of sparsity on accuracy w/ 100% shift (e) Impact of iterative training on performance w/o shift (f) Impact of iterative training on performance w/ 100% shift

Analysis masks

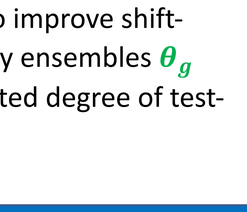
Personalized masks



divide 20 clients into 5 groups, with similar data distribution within groups

- the similarity of masks is correlated with local training data distribution

Shared weights



15 clients' local data are similar to global distribution

5 clients' local data are heterogeneous

- the ratio of shared weights is correlated with data heterogeneity

Acknowledgment

We are grateful to anonymous reviewers for their constructive comments. This work is partially supported by National Science Foundation of China (NSFC) under Grant No. U21A20516 and 62076017, the Beihang University Basic Research Funding No. YWF-22-L-531, the Funding No. 22-TQ23-14-ZD-01-001, the CityU APRC grant No. 9610633, and WeBank Scholars Program. Yongxin Tong is the corresponding author.