Adaptive Test-Time Personalization for Federated Learning

Wenxuan Bao [1]* Tianxin Wei [1]* Haohan Wang [1] Jingrui He [1]

[1]University of Illinois Urbana-Champaign

AIFARMS Artificial Intelligence for Future Agricultural Resilience, Management, and Sustainability

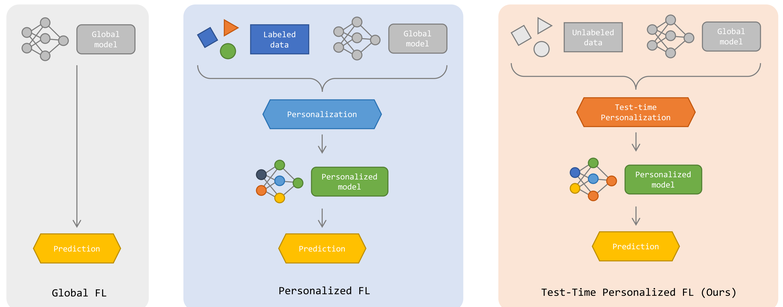
Background

Cross-device federated learning A very large number of mobile/IoT devices collaborate to train machine learning models under the orchestration of a central server, without sharing their raw data.

- Only a small part of clients have labeled data and are sampled for training.
- However, the model also needs to be deployed on clients that do not participate in FL training.
- Clients typically have their own distributions with distribution shifts, e.g., feature shift, label shift.

Question: How to generalize to unparticipating clients under distribution shifts?

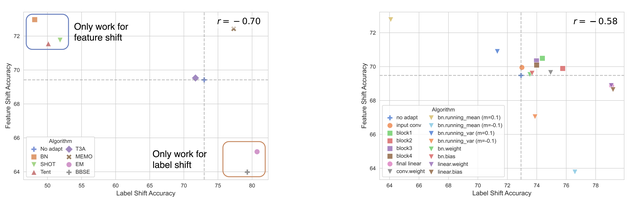
Test-Time Personalized Federated Learning (TTPFL)



Framework	Global FL	Personalized FL	Test-Time Personalized FL
Adaptation to each client	No :(	Yes :)	Yes :)
Data requirement	No :)	Additional labeled data :(	Unlabeled testing data :)

Test-time adaptation (TTA) methods can be applied to TTPFL, with **two drawbacks**:

- *Neglection of multiple sources*: TTA assumes single source domain and neglects the interrelationship among source clients, resulting in sub-optimal generalization.
- *Inflexibity to various distribution shifts*. Most TTA methods are customized for specific distribution shifts and lack the flexibility to address diverse types of distribution shifts.



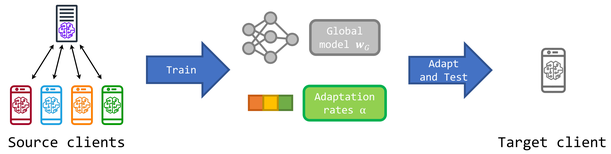
Motivation: Which modules to adapt should depend on the type of distribution shifts among clients, which can be inferred from source clients.

Paper Summary

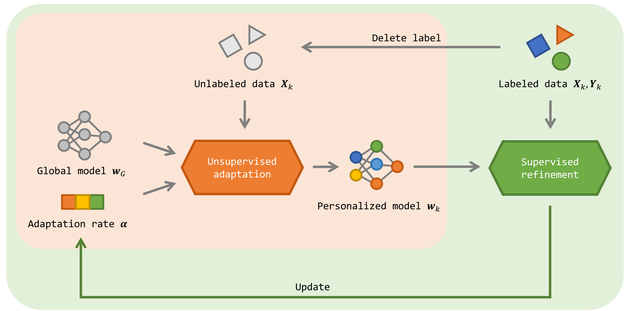
- We consider a novel setting named *Test-Time Personalized Federated Learning*, addressing the challenge of personalizing a global model to each *unparticipating client* during test-time, without requiring any labeled data.
- We propose *ATP*, which adaptively learns the adaptation rate for each module, enabling it to handle different types of distribution shifts among FL clients.

Proposed Method: ATP

Adaptive test-time personalization (ATP) learns the adaptation rates for each module.



Training phase: Learn to adapt with source clients



Unsupervised adaptation: Simulate adapting the global model with *unlabeled data*.

$$\boldsymbol{w}_k \leftarrow \boldsymbol{w}_G + (\boldsymbol{A}\boldsymbol{\alpha}) \odot \boldsymbol{h}_k \quad \text{i.e.,} \quad \boldsymbol{w}_k^{[l]} \leftarrow \boldsymbol{w}_G^{[l]} + \alpha^{[l]} \boldsymbol{h}_k^{[l]} \text{ for each module } l, \text{ where}$$

- For trainable parameters, $\boldsymbol{h}_k^{[l]} = -\nabla_{\boldsymbol{w}^{[l]}} \ell_H(f(\boldsymbol{X}_k; \boldsymbol{w}_G))$ (negative gradient of entropy)
- For BatchNorm running statistics, $\boldsymbol{h}_k^{[l]} = \hat{\boldsymbol{w}}_k^{[l]} - \boldsymbol{w}_G^{[l]}$ (batch stat $-$ global stat)

Supervised refinement: Refine the adaptation rates $\boldsymbol{\alpha}$ with *labeled data* and cross entropy

$$\boldsymbol{\alpha} \leftarrow \boldsymbol{\alpha} - \eta \nabla_{\boldsymbol{\alpha}} \ell_{CE}(f(\boldsymbol{X}_k; \boldsymbol{w}_k), \boldsymbol{Y}_k)$$

Server aggregation: Local adaptation rates are uploaded to and aggregated by the server.

Testing phase: Exploit adaptation rates on target clients

ATP-batch processes each batch independently. It is identical to unsupervised adaptation.

ATP-online processes a stream of batches $[\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots]$. It uses update directions on previous batches $[\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_k]$ to assist the adaptation to the current batch.

$$\boldsymbol{w}_k \leftarrow \boldsymbol{w}_G + (\boldsymbol{A}\boldsymbol{\alpha}) \odot \left(\frac{1}{k} \sum_{s=1}^{k} \boldsymbol{h}_s \right)$$

Experiments

ATP can handle different types of distribution shifts

We train ResNet-18 on CIFAR-10(C) with different types of distribution shifts:

- Feature shift: Each client has a random type of image corruption.
- Label shift: Each client has 2 majority classes and 8 minority classes.
- Hybrid shift: Feature + label shifts.

Table 1:Accuracy (mean $\pm$ s.d. %) on target clients under various distribution shifts on CIFAR-10(C)

Method	Feature shift	Label shift	Hybrid shift	Avg. Rank
No adaptation	69.42 ± 0.13	72.98 ± 0.24	63.68 ± 0.24	7.7
BN-Adapt	73.52 ± 0.22	54.54 ± 0.10	50.42 ± 0.39	7.0
SHOT	71.76 ± 0.17	48.13 ± 0.18	44.68 ± 0.32	9.3
Tent	71.76 ± 0.09	50.13 ± 0.21	46.05 ± 0.26	8.3
T3A	69.53 ± 0.08	71.70 ± 0.32	62.17 ± 0.17	8.0
MEMO	72.43 ± 0.22	77.30 ± 0.15	68.07 ± 0.28	4.3
EM	65.18 ± 0.12	80.73 ± 0.18	69.85 ± 0.43	5.0
BBSE	63.98 ± 0.17	79.30 ± 0.17	67.96 ± 0.43	6.7
Surgical	69.85 ± 0.22	76.00 ± 0.17	66.94 ± 0.43	6.3
ATP-batch	73.68 ± 0.10	79.90 ± 0.22	73.05 ± 0.35	2.3
ATP-online	**74.06 ± 0.18**	**81.96 ± 0.14**	**75.37 ± 0.22**	**1.0**

ATP consistently improves the performance across different types of distribution shifts. We also conduct experiments on more datasets (CIFAR-100(C), Digits-5, PACS), and try different models (Shallow-CNN, ResNet-18, ResNet-50).

ATP learns shift-specific adaptation rates

Table 2:Train and test adaptation rates with different distribution shifts, accuracy (mean $\pm$ s.d. %)

Train	Test		
	Feature shift	Label shift	Hybrid shift
No adaptation	69.42 ± 0.13	72.98 ± 0.24	63.68 ± 0.24
Feature shift	**73.68 ± 0.10**	65.05 ± 1.82	60.64 ± 1.43
Label shift	67.99 ± 0.28	**79.90 ± 0.22**	69.50 ± 0.52
Hybrid shift	72.69 ± 0.14	78.92 ± 0.34	**73.05 ± 0.35**

- ATP performs the best when training and testing under the same type of distribution shift.
- The adaptation rates trained under feature shifts have negative impact on label shifts, and vice versa.
- The adaptation rates trained under hybrid shift are also beneficial for feature and label shifts.

Acknowledgement This work is supported by National Science Foundation under Award No. IIS-1947203, IIS-2117902, IIS-2137468, IIS-2002540, Agriculture and Food Research Initiative (AFRI) grant no. 2020-67021-32799/project accession no.1024178 from the USDA National Institute of Food and Agriculture, the U.S. Department of Homeland Security under Grant Award Number, 17STQAC00001-06-00, and IBM-Illinois Discovery Accelerator Institute - a new model of an academic-industry partnership designed to increase access to technology education and skill development to spur breakthroughs in emerging areas of technology. The views and conclusions are those of the authors and should not be interpreted as representing the official policies of the funding agencies or the government.