PerAda: Parameter-Efficient Federated Learning Personalization with Generalization Guarantees

Chulin Xie[1] De-An Huang[2] Wenda Chu[3] Daguang Xu[2] Chaowei Xiao[2,4] Bo Li[1,5] Anima Anandkumar[3]

[1] University of Illinois Urbana-Champaign [2] NVIDIA [3] Caltech [4] University of Wisconsin-Madison [5] University of Chicago

https://github.com/NVlabs/PerAda

CVPR SEATTLE, WA JUNE 17-21, 2024

Background & Motivation

Personalized Federated Learning (FL)

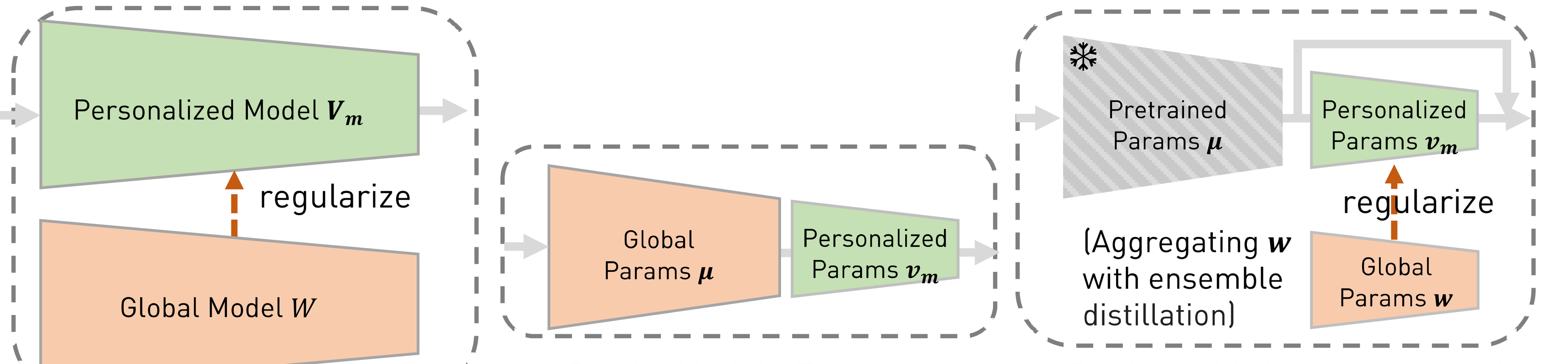
- Personalization: each client trains a personalized model on local data
- Generalization: clients leverage aggregated knowledge from other clients

Existing work

- Full model personalization: Each client trains a personalized model and a copy of global model aggregated by the server for regularization
- Partial model personalization:
  - Split model into personalized and shared parameters
  - Only shared parameters are aggregated

Challenge

- High communication and computational costs: Full model personalization doubles the memory at each client
- Limited generalization: Partial model personalization overfits to local data and struggles with distribution shifts
  - shared parameters do not encode generalized knowledge well compared to full global model

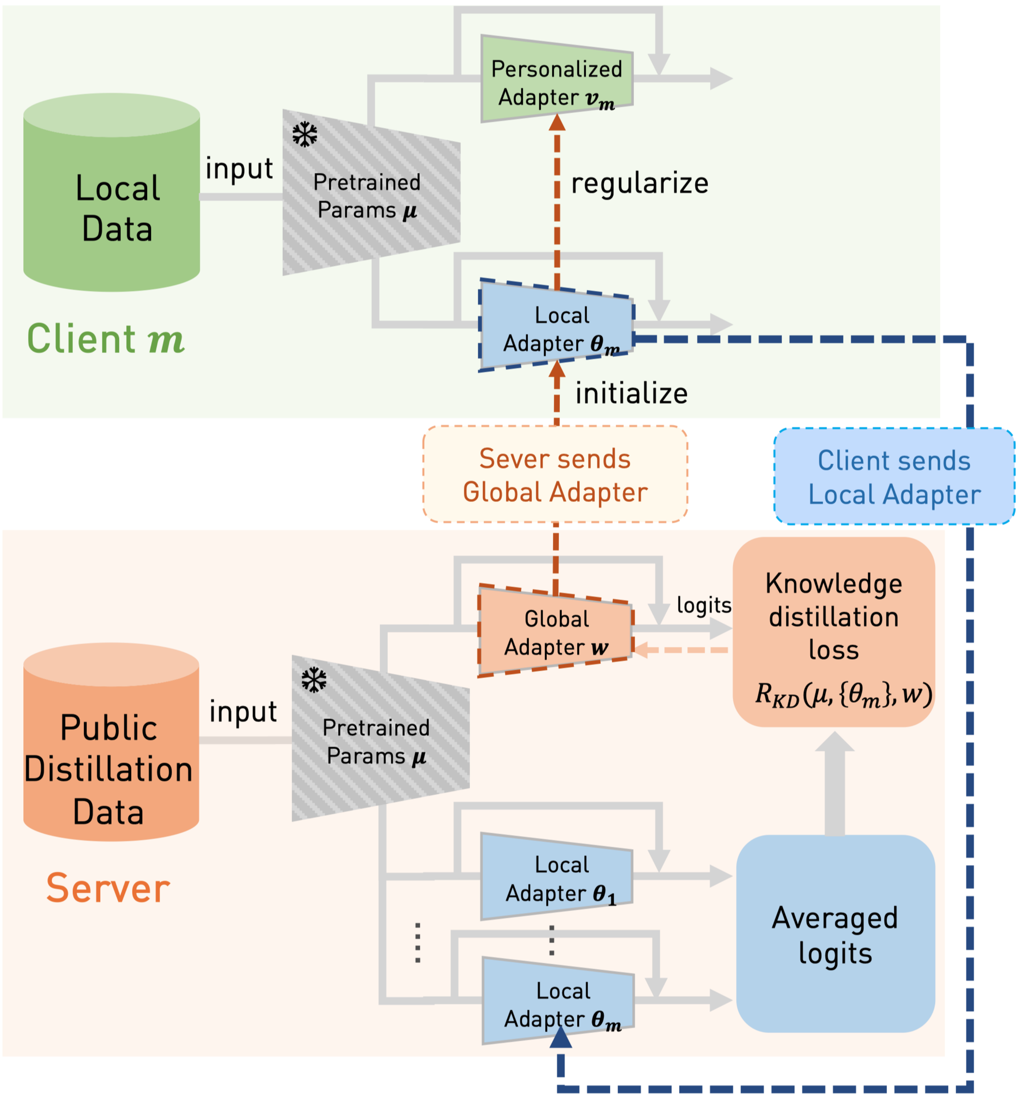


(a) Full Model pFL (b) Partial Model pFL (e.g., personalized output layer) (c) PerAda with frozen ❄ pretrained model and personalized adapter

This Work: PerAda

- Framework: parameter-efficient personalized FL using Adapters and Knowledge Distillation (KD)
- Benefits:
  - Reduce communication and computation costs with a pretrained model and adapters
  - Achieve personalization while maintaining generalization to test-time distribution shifts with regularization and KD
- Theoretical justification of PerAda (in paper)
  - Convergence analysis for global & personalized models
  - generalization guarantees for global & personalized models

Method



Personalized objective:

- train personalized adapter with regularization towards a global adapter to prevent overfitting

$$\min_{v_m} P_m(v_m, w) := \mathcal{L}_m(u, v_m) + \frac{\lambda}{2}\|v_m - w\|^2, \quad \text{(Personal Obj)}$$

Global objective:

- leverage server-side ensemble distillation to enrich the global adapter with ensemble knowledge from clients' local models
- avoid directly averaging clients' models

$$\min_{w} \mathcal{R}_{\mathrm{KD}}(u, \{\theta_m\}_{m=1}^{M}, w) \quad \text{(Global Obj)}$$

where $\theta_m = \arg\min_{\theta} \mathcal{L}_m(u, \theta)$, initialized with w.

$\mathcal{R}_{\mathrm{KD}}$ is average distillation loss (between the averaged logits of local models and logits of the global model) on an auxiliary (unlabeled) dataset

Experiments

- Local-test: clients' corresponding local test data → personalization
- Global-test: the union of clients' local test data → generalization

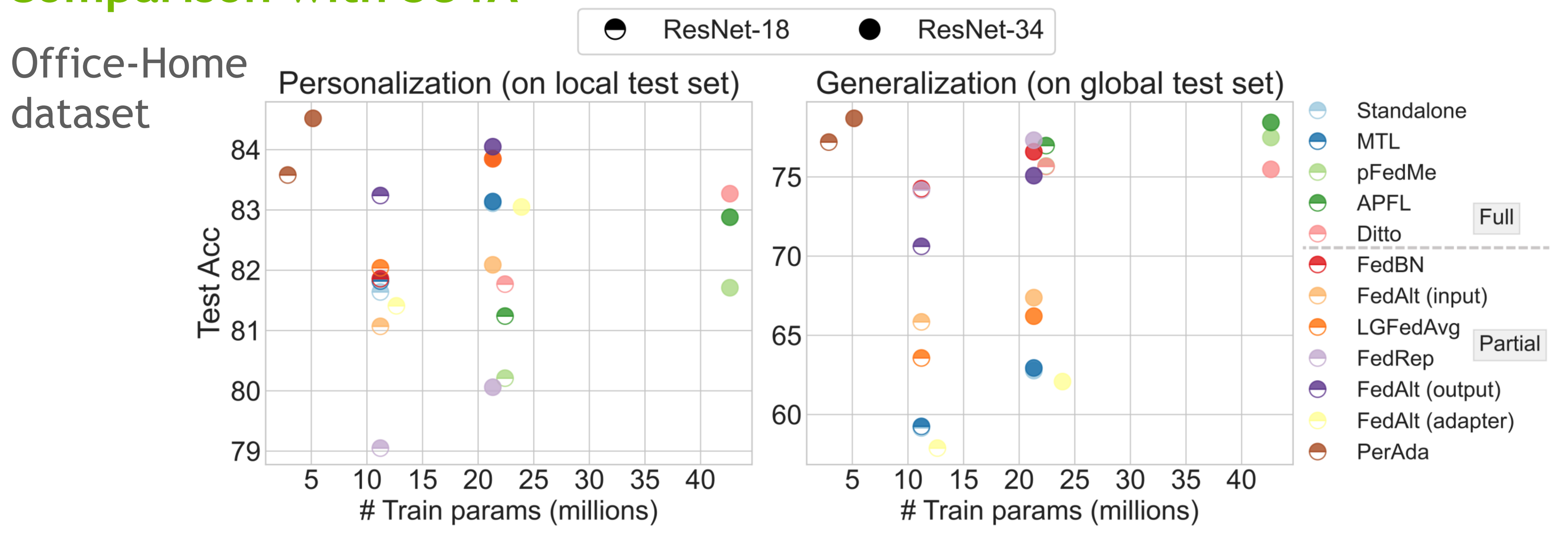
Parameter-efficiency and averaged test accuracy across all clients' personalized models

Algorithm	Personalized Params	# Trained Params	# Comm. Params	CIFAR-10 Local-test	CIFAR-10 Global-test	CIFAR-10.1	CIFAR-10-C	Office-Home Local-test	Office-Home Global-test	CheXpert Local-test	CheXpert Global-test
STANDALONE	Full model	11.18M	**0**M	85.94± 8.82	29.77± 8.09	25.82± 6.27	26.67 ± 7.07	81.64± 6.08	59.15± 3.32	65.06± 1.88	65.45± 2.3
MTL [57]	Full model	11.18M	11.18M	86.24± 8.45	29.46± 8.33	25.64± 6.42	26.4 ± 7.29	81.82± 5.53	59.25± 2.84	65.15± 1.95	65.48± 2.3
FEDAVG+FT [65]	Full model	11.18M	11.18M*	88.91± 5.71	43.99± 9.57	35.49± 8.02	36.51 ± 8.36	79.42± 5.62	<u>77.19</u>± 0.56	70.16± 0.78	70.6 ± 0.31
PFEDME [59]	Full model	22.36M	11.18M	90.73± 4.67	45.06± 8.65	36.51± 7.2	37.65 ± 7.6	80.21± 5.32	75.69± 0.69	65.07± 1.2	64.86± 1.22
APFL [10]	Full model	22.36M	11.18M	90.74± 4.75	43.92± 9.18	35.83± 7.5	36.51 ± 7.94	81.24± 4.51	76.98± 1.39	68.98± 1.04	68.96± 1.1
DITTO [33]	Full model	22.36M	11.18M	90.21± 4.61	<u>53.82</u>± 6.35	<u>42.72</u>± 5.68	44.32 ± 5.73	81.77± 4.31	75.66± 1.01	68.79± 1.4	68.86± 1.22
FEDBN [36]	Batch norm.	11.18M	11.17M	90.37± 5.19	43.18± 8.67	35.01± 7.24	36.29 ± 7.43	81.86± 5.13	74.26± 0.52	68.74± 1.17	68.83± 1.08
FEDALT [48]	Input layer	11.18M	6.45M	87.07± 6.54	32.23± 8.23	27.49± 6.41	28.51 ± 7.11	81.07± 5.59	65.85± 0.9	67.63± 1.18	67.74± 1.1
FEDSIM [48]	Input layer	11.18M	6.45M	87.93± 6.25	33.07± 8.16	28.21± 6.41	29.15 ± 7.16	82.45± 5.03	67.66± 0.82	67.49± 1.32	67.54± 1.24
LG-FEDAVG [38]	Feat. extractor	11.18M	0.005M	86.7 ± 8.01	29.96± 8	25.97± 6.21	26.83 ± 6.95	82.04± 5.96	63.57± 2.32	65.78± 1.62	66.23± 1.75
FEDREP [9]	Output layer	11.18M	11.17M	87.76± 6.46	35.19± 6.97	30.15± 5.89	30.68 ± 6.31	79.05± 5.88	74.17± 2.02	66.66± 1.82	66.52± 1.47
FEDALT [48]	Output layer	11.18M	11.17M	89.68± 5.4	40.68± 7.3	33.61± 6.12	34.3 ± 6.5	83.24± 3.96	70.62± 1.46	68.27± 1.3	68.36± 1.31
FEDSIM [48]	Output layer	11.18M	11.17M	89.75± 5.51	41.98± 7.66	34.21± 6.22	35.31 ± 6.79	82.91± 4.46	72.34± 0.51	68.22± 1.34	68.12± 1.24
FEDALT [48]	Adapter	12.59M	11.18M	87.26± 7.78	31.51± 8.55	27.38± 6.65	27.77 ± 7.19	81.41± 6.5	57.88± 3.57	72.13± 1.34	74.67± 1.57
FEDSIM [48]	Adapter	12.59M	11.18M	87.76± 7.57	31.97± 7.44	27.76± 5.78	28.1 ± 6.46	82.14± 5.46	58.62± 3.24	71.75± 1.4	74.09± 1.55
PERADA W/O KD	Adapter	2.82M	1.41M	<u>91.27</u>± 5.15	53.81± 6.27	42.5 ± 5.06	<u>44.45</u>± 5.48	<u>83.31</u>± 5.54	76.55± 2.47	<u>76.77</u>± 2.24	<u>77.59</u>± 2.18
PERADA	Adapter	**2.82**M	1.41M	**91.82**± 4.43	**59.05**± 5.24	**47.25**± 4.48	**48.53**± 4.74	**83.58**± 4.74	**77.2** ± 1.63	**76.98**± 3.87	**77.88**± 1.55

- PerAda achieves the highest personalized performance and generalization by updating the smallest number of model parameters
- Existing partial model personalization methods have poor generalization to distribution shifts.
- Adapter-based personalization methods are generally effective on CheXpert.

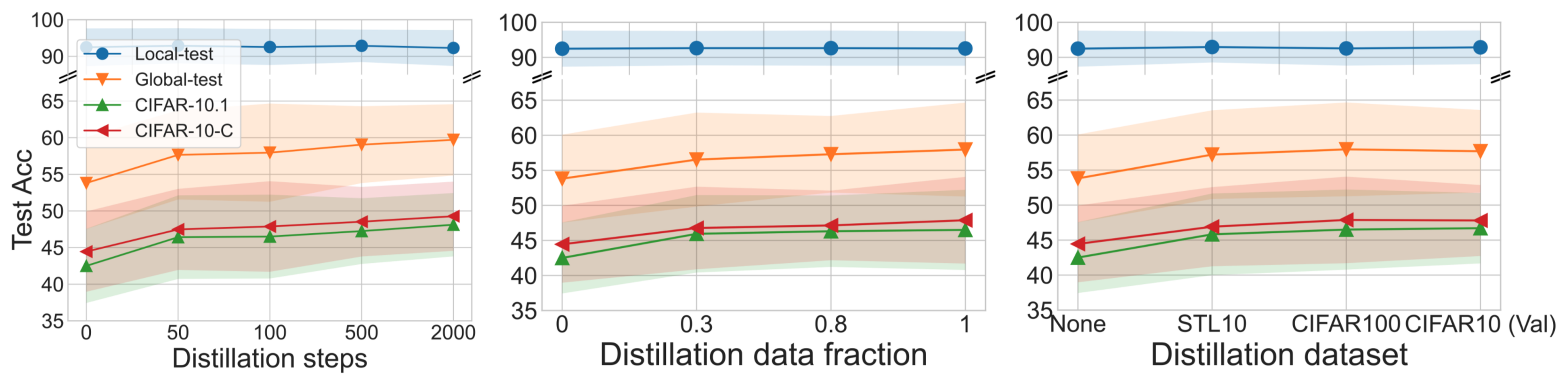
Comparison with SOTA

Office-Home dataset



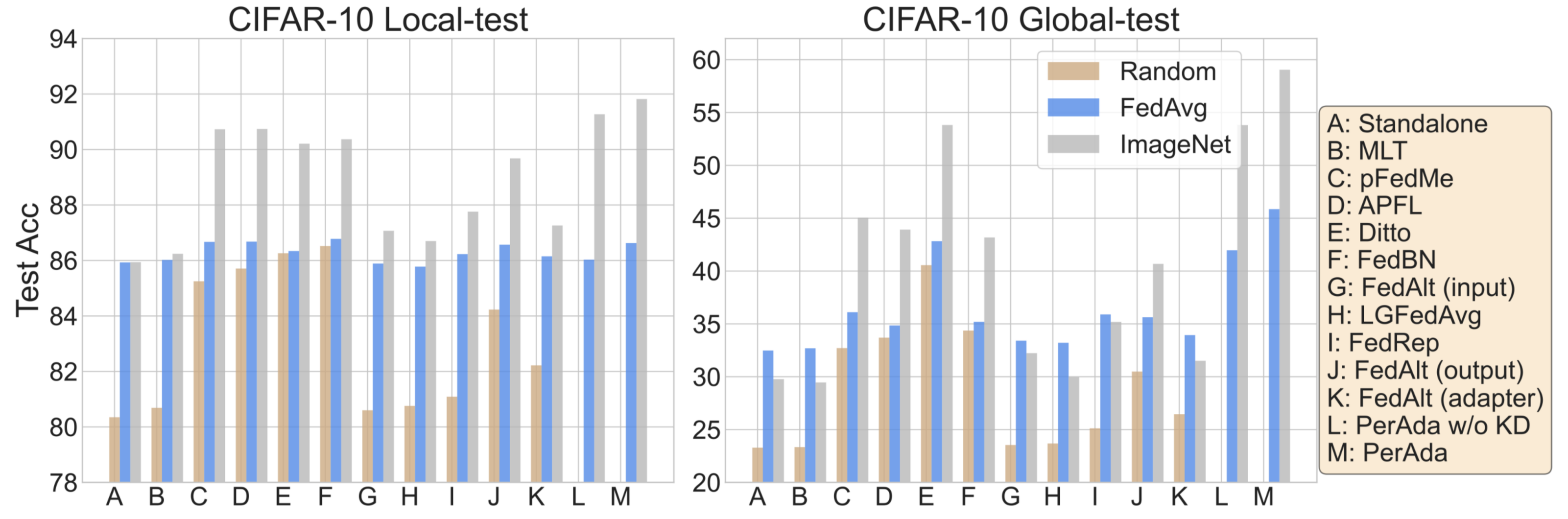
Effect of knowledge distillation

CIFAR-10 dataset



- More distillation steps & data help
- Out-of-domain distillation data achieve similar performance as in-domain data

Effect of pretrained models



- ImageNet-pretraining leads to better performance than Fed-Avg pretraining

Utility under differential privacy (DP) guarantees

CIFAR-10 dataset with ViT-S/16-224 model

Algorithm	Personalization	$\epsilon = \infty$	$\epsilon = 5.99 \pm 3.03$	$\epsilon = 3.7 \pm 2.12$	$\epsilon = 1.81 \pm 1.12$
Ditto	Full	**98.59** ± 1.63	76.76 ± 24.14	76.75 ± 24.13	76.67 ± 24.12
PERADA W/O KD	Adapter	97.69 ± 1.79	77.49 ± 21.21	77.32 ± 21.16	76.68 ± 21
PERADA	Adapter	98.08 ± 1.28	**80.33**± 20.76	**79.79** ±20.45	**77.83**± 19.58

- Perform local training with DP-SGD for personalized & global model
- PerAda achieves higher utility than full model personalization under DP