Flexible Sharpness-Aware Personalized Federated Learning

AAAI-25 / IAAI-25 / EAAI-25
FEBRUARY 25 – MARCH 4, 2025 | PHILADELPHIA, USA

Xinda Xing[1,2*], Qiugang Zhan[3,4,5*], Xiurui Xie[1†], Yuning Yang[2], Qiang Wang[6], Guisong Liu[3,4,5†]

1. Laboratory of Intelligent Collaborative Computing, UESTC *2. Computer Science and Engineering, UESTC*
3. Complex Laboratory of New Finance and Economics, SWUFE *4. Engineering Research Center of Intelligent Finance, Ministry of Education, PRC*
5. Kash Institute of Electronics and Information Industry *6. School of Cyber Science and Technology, SYSU*

Introduction and Motivation

Introduction

Personalized Federated Learning (**PFL**) is a new paradigm to address the statistical heterogeneity problem in federated learning. Unlike previous PFL methods, our **FedFSA** customizes a personalized sharpness-aware minimization (**SAM**) optimizer based on the client's local loss characteristics.

Motivation

SAM introduces a perturbation during gradient ascent.

1. DISAM and FedLESAM suggest that **larger perturbations improve generalization but slow convergence**.
2. SSAM and SAMON show that **not all parameters** need perturbation**.**

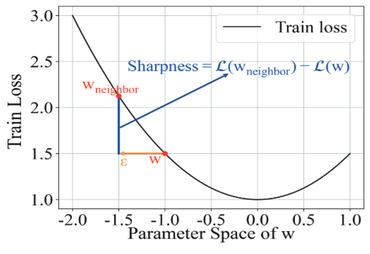
So we hypothesize slow convergence may arise from applying large perturbations to unnecessary layers.

Contributions: Our FedFSA can flexibly select the layers with the highest sharpness to employ larger perturbation, achieving better performance and converges more quickly.

Sharpness-Aware Minimization

What is Sharpness?

$$\max_{\|\boldsymbol{\epsilon}\|_2 \le \rho} \mathcal{L}(W+\boldsymbol{\varepsilon}) - \mathcal{L}(W)$$



How steep or flat the losslandcape is around a neighborhood, which can affect the generalization ability of a model.

Sharpness-Aware Minimization

The objective of SAM is to minimize the loss function and smooth the loss landscape by solving the following min-max problem:

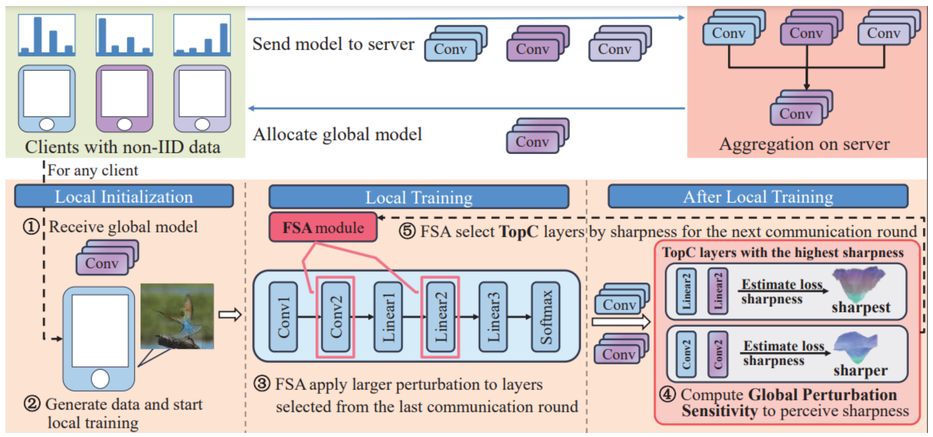
$$\min_{W} \max_{\|\boldsymbol{\epsilon}\|_2 \le \rho} \mathcal{L}(W+\boldsymbol{\varepsilon})$$

SAM optimization in **2 Steps:**

1. **Gradient Ascent with Perturbation**: Adjust parameters to increase loss by adding a perturbation influenced by the perturbation amplitude ρ.
2. **Gradient Descent**: Update parameters to reduce loss based on the new gradient.

Methodology

FedFSA Workflow



Perturbation Sensitivity to Perceive Sharpness

Sharpness Difinition: $\max_{\|\boldsymbol{\epsilon}\|_2 \le \rho} \mathcal{L}(W+\boldsymbol{\varepsilon}) - \mathcal{L}(W)$

Perturbation Sensitivity: The change in model output or loss after removing the perturbation of the k-th layer parameters.

$$\begin{aligned} s_k &= |\mathcal{L}(W+\boldsymbol{\varepsilon}) - \mathcal{L}(w_1+\epsilon_1, \ldots, w_k, \ldots, w_L+\epsilon_L)| \\ &= |\nabla_{w_k}\mathcal{L}(W+\boldsymbol{\varepsilon})\cdot\epsilon_k + R_1(W+\boldsymbol{\varepsilon})| \\ &\approx |\nabla_{w_k}\mathcal{L}(W+\boldsymbol{\varepsilon})\cdot\epsilon_k|. \end{aligned}$$

Global Perturbation Sensitivity: The variation in parameters during training can be considered as a large perturbation, added to the local model to reasonably explore the neighborhood. In communication round t, for client i, we have:

$$\begin{aligned} s^t_{i,k} &\approx |\Delta w_{i,k}\cdot\epsilon_{i,k}| \\ &= |(w^{t,E}_{i,k} - w^{t,0}_{i,k})\cdot\epsilon_{i,k}| \\ &= (w^{t,E}_{i,k} - w^{t,0}_{i,k})^2. \end{aligned}$$

Experimental Validation

Comparison with Baselines

Method	FMNIST(%)		CIFAR10(%)		CIFAR100(%)		TINY(%)	
	Pat(5)	Dir(0.5)	Pat(5)	Dir(0.5)	Pat(15)	Dir(0.3)	Pat(50)	Dir(0.3)
FedAvg-FT	92.51±0.21	91.59±0.47	81.79±0.61	81.41±0.77	53.70±0.26	41.09±0.14	23.03±0.46	20.75±0.35
FedCR	93.67±0.11	92.43±0.41	84.00±0.14	83.49±0.32	59.91±0.35	42.74±0.39	23.71±0.78	—
FedALA	92.99±0.22	91.83±0.39	81.62±1.08	81.76±0.53	54.92±0.20	39.75±1.07	—	—
FedSAM	93.01±0.11	91.89±0.39	84.16±0.20	83.84±0.44	59.41±0.39	44.66±0.40	28.98±0.15	25.55±0.53
MoFedSAM	**93.79±0.21**	**92.94±0.27**	88.33±0.14	88.16±0.25	70.33±0.29	51.02±2.21	33.80±0.29	28.02±0.60
FedSMOO	93.51±0.05	92.46±0.19	87.22±0.16	87.03±0.24	66.56±0.68	52.34±1.17	26.29±0.02	21.96±0.04
FedSpeed	93.68±0.15	92.88±0.20	88.24±0.37	87.99±0.25	68.24±0.15	52.61±0.55	29.60±0.28	24.92±0.43
FedFSA	93.78±0.11	92.88±0.35	**88.64±0.09**	**88.37±0.27**	**72.28±0.40**	**60.87±0.61**	**41.39±0.25**	**33.76±0.93**

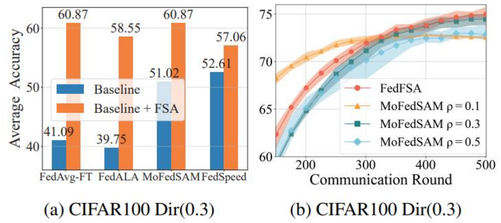
Table 1: Average test accuracy under Pathological and Dirichlet non-IID settings on FMNIST, CIFAR10, CIFAR100, and TINY. Bold fonts highlight the best accuracy.

Impact of Heterogeneity and Scalability

Method	Heterogenity		Scalability	
	CIFAR100(%)		TINY(%)	
	Dir(0.1)	Dir(1)	50 clients	100 clients
FedAvg-FT	52.92±0.46	34.58±0.50	22.24±0.39	20.75±0.35
FedCR	55.87±0.44	30.31±0.36	28.62±0.39	—
FedALA	52.56±1.32	34.72±0.63	—	—
FedSAM	58.42±1.08	37.40±0.24	26.73±0.12	25.55±0.53
MoFedSAM	66.79±0.75	44.41±0.59	28.56±0.38	28.02±0.60
FedSMOO	65.85±0.71	46.58±0.52	25.87±0.27	21.96±0.04
FedSpeed	65.98±0.91	46.19±0.56	31.60±0.25	24.92±0.43
FedFSA	**67.73±0.69**	**52.49±0.35**	**36.90±0.43**	**33.76±0.93**

Table 2: Average test accuracy at different levels of heterogeneity on CIFAR100 and scalability with different numbers of clients on TINY.

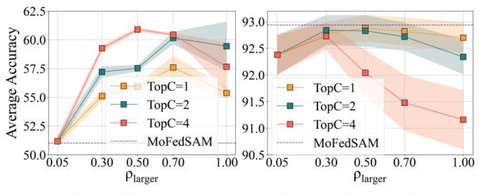
Applicability Evaluation



(a) CIFAR100 Dir(0.3) (b) CIFAR100 Dir(0.3)

Figure 3: Average test accuracy to demonstrate the applicability of FSA. (a) shows the applicability of FSA to other FL methods and (b) shows the applicability of FSA to ResNet18.

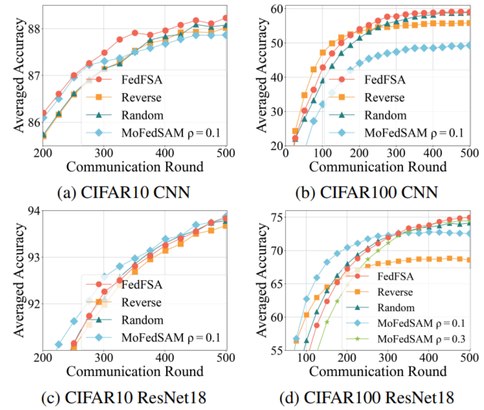
Hyperparameter Experiment



(a) CIFAR100 Dir(0.3) (b) FMNIST Dir(0.5)

Figure 4: The impact of the hyperparameter ρ_{larger} and TopC of FedFSA on different types of datasets, ranging from complex to simple. The perturbation amplitude ρ_{default} for MoFedSAM is set to 0.1, while for FedFSA is 0.05.

Ablation Evaluation



(a) CIFAR10 CNN (b) CIFAR100 CNN

(c) CIFAR10 ResNet18 (d) CIFAR100 ResNet18

Figure 5: The effect of different critical parameter selection schemes on CIFAR10 and CIFAR100.