Combating Exacerbated Heterogeneity for Robust Models in Federated Learning

Jianing Zhu[1], Jiangchao Yao[2,3], Tongliang Liu[4], Quanming Yao[5], Jianliang Xu[1], Bo Han[1]

[1]Hong Kong Baptist University [2]Shanghai Jiao Tong University [3]Shanghai AI Laboratory

[4]Sydney AI Centre, The University of Sydney [5]Tsinghua University

Outline

- Background
- Intensified Heterogeneity
- Slacked Federated Adversarial Training (SFAT)
- Summary

SFAT: Slack Federated Adversarial Training

arXiv 2303.00250 | OpenReview SFAT | Github | Pub ICLR'23 | Slides SFAT

This repo contains the sample code of our proposed framework `Slack Federated Adversarial Training (SFAT)` in our paper: Combating Exacerbated Heterogeneity for Robust Models in Federated Learning (ICLR 2023).

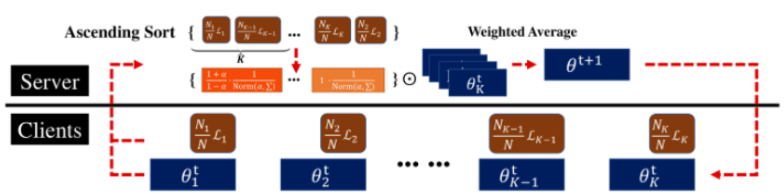
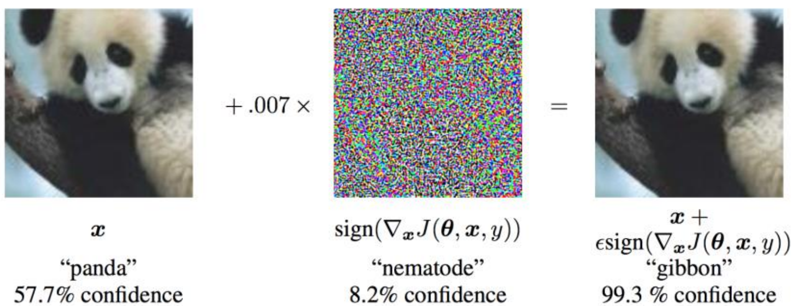


Figure. Framework overview of SFAT.

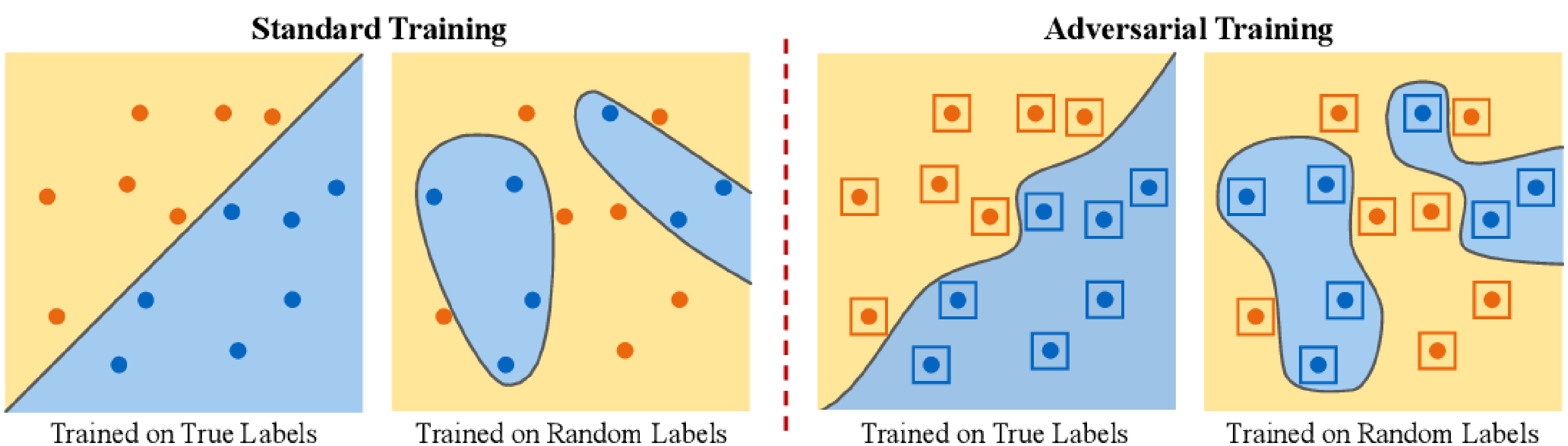
Adversarial Training

Vulnerability of Deep Neural Networks (DNNs)



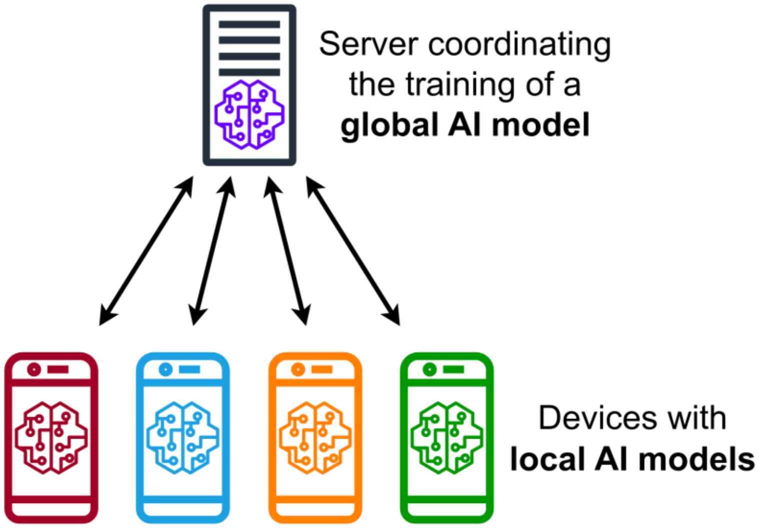
DNNs could be easily fooled by adding a small number of perturbations on natural inputs. The perturbed example is adversarial example.

$$\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} \max_{\|\mathbf{x}_i' - \mathbf{x}_i\|_p \leq \epsilon} \ell(h_{\boldsymbol{\theta}}(\mathbf{x}_i'), y_i)$$



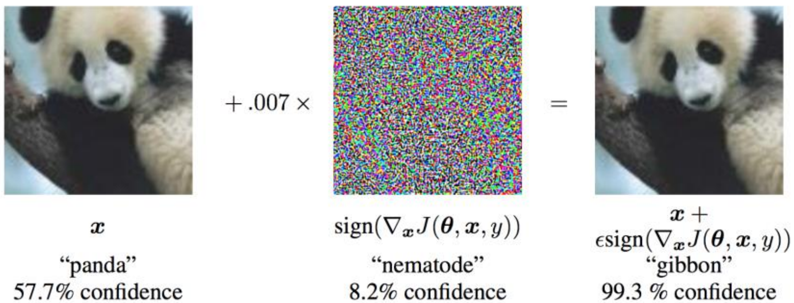
Federated Adversarial Training

Adversarial Vulnerability in Device-edge



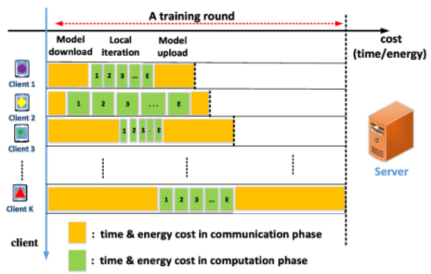
a. Local Adversarial Training

b. Federated Model Aggregation

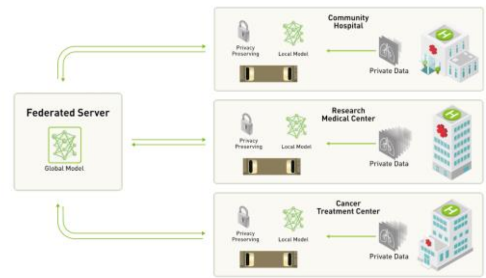


Previous Work

*Towards **distributed adaptability** of federated adversarial training*



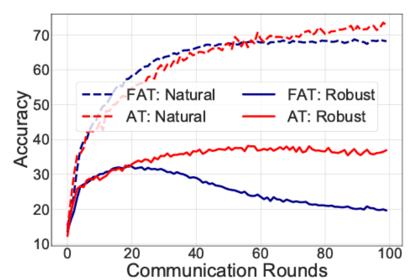
Communication Cost



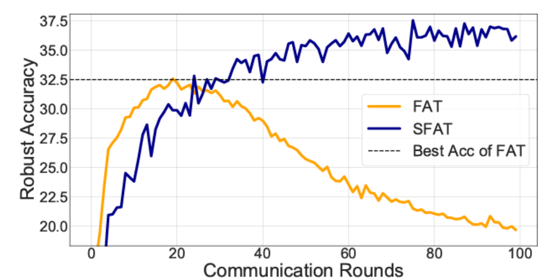
Hardware Capacity

Issues on Distributed Systems

Unexpected Robust Deterioration



(a) Centralized AT vs. FAT

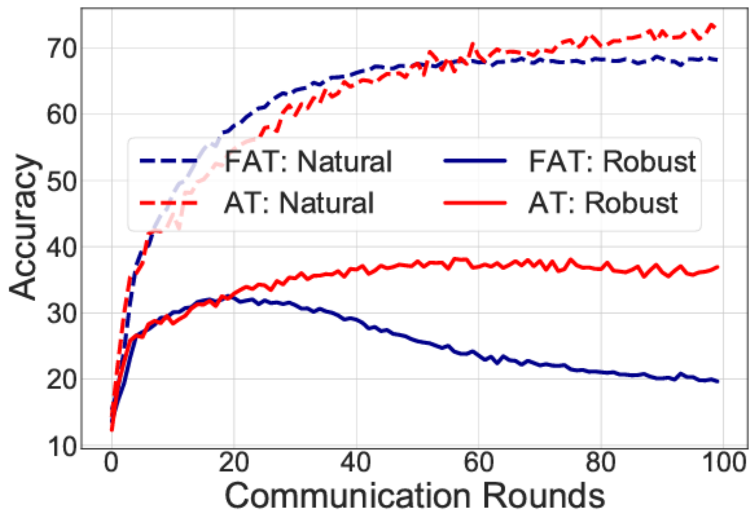


(b) FAT vs. SFAT (ours)

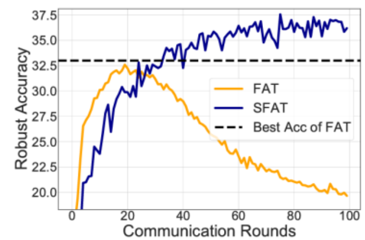
*Towards the **algorithmic challenge** of federated adversarial training*

Unexpected Robust Deterioration

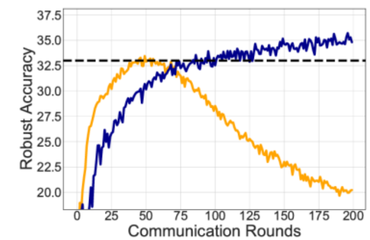
AT vs. FAT



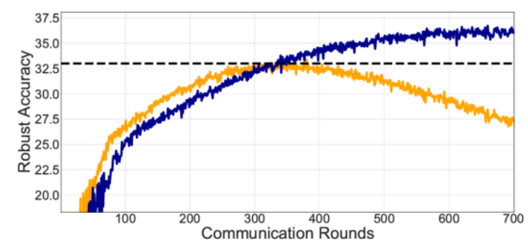
(a) Centralized AT vs. FAT



(a) Local training epoch: 10



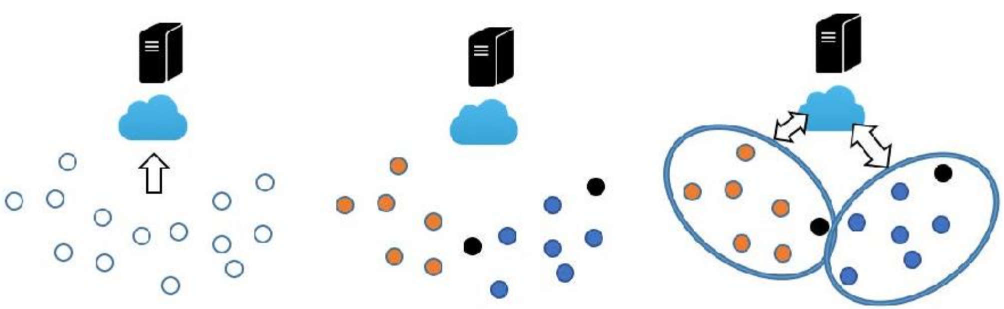
(b) Local training epoch: 5



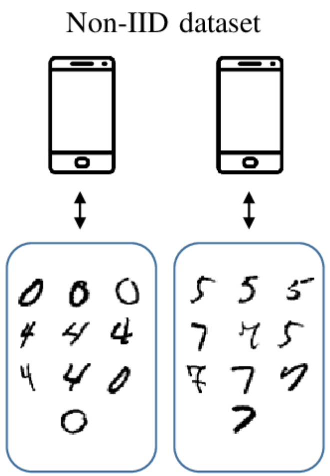
(c) Local training epoch: 1

Intensified Heterogeneity

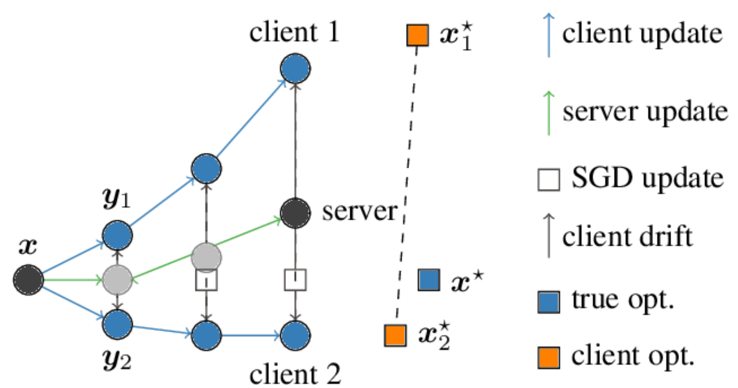
Heterogeneity in FL



(i) (ii) (iii)

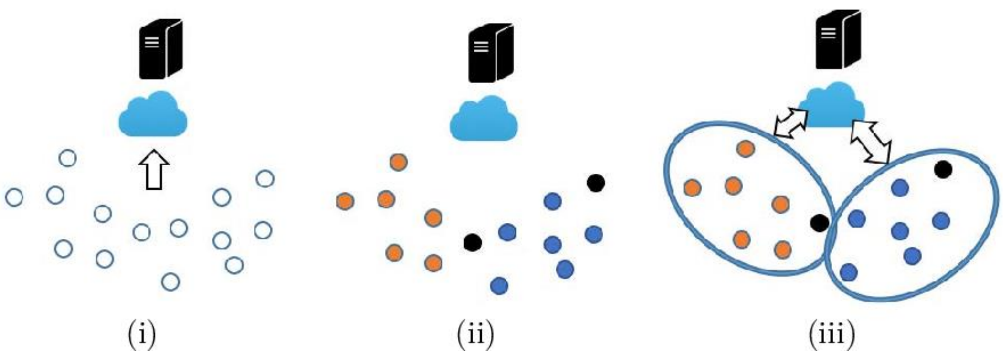


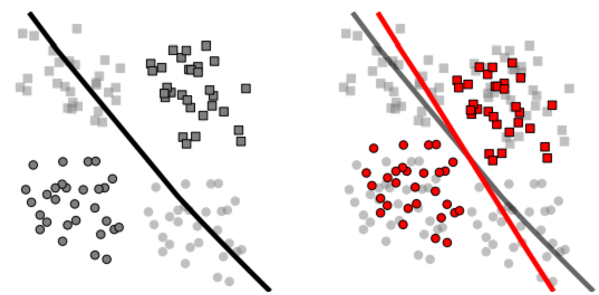
$$\min_{f_\theta \in \mathcal{F}} \frac{1}{N} \sum_{n=1}^{N} \ell(f_\theta(x_n), y_n)$$

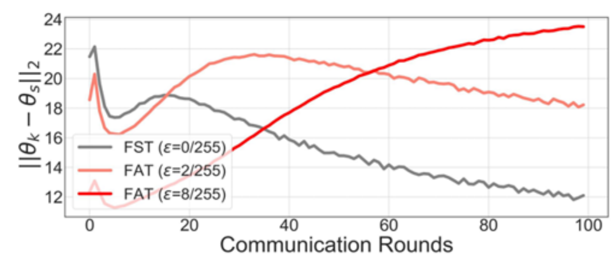


Intensified Heterogeneity

Dynamic Heterogeneity



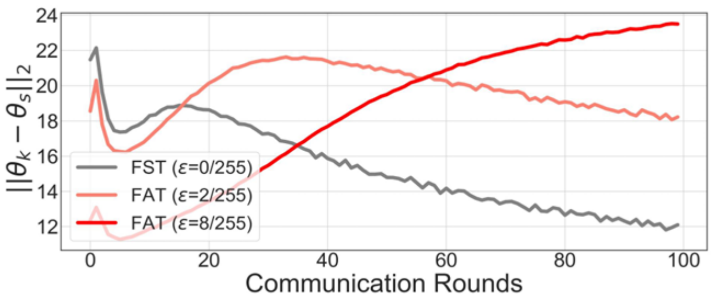




$$\min_{f_\theta \in \mathcal{F}} \frac{1}{N} \sum_{n=1}^{N} \boxed{\max_{\tilde{x}_n \in \mathcal{B}_\epsilon[x_n]} \ell(f_\theta(\tilde{x}_n), y_n)}$$

Intensified Heterogeneity

Adversarial Training with FL



$$\min_{f_\theta \in \mathcal{F}} \frac{1}{N} \sum_{n=1}^{N} \boxed{\max_{\tilde{x}_n \in \mathcal{B}_\epsilon[x_n]} \ell(f_\theta(\tilde{x}_n), y_n)}$$

ε ↑ client drift ↑

the inner-maximization for pursuing adversarial robustness would exacerbate the data heterogeneity among local clients in federated learning.

Intensified Heterogeneity

α-Slack Mechanism

$$\mathcal{L}_{AT} = \frac{1}{N}\sum_{n=1}^{N}\max_{\tilde{x}_n \in \mathcal{B}_\epsilon[x_n]} \ell(f_\theta(\tilde{x}_n), y_n)$$

$$\mathcal{L}_{AT} = \frac{1}{N}\sum_{n=1}^{N}\max_{\tilde{x}_n \in \mathcal{B}_\epsilon[x_n]} \ell(f_\theta(\tilde{x}_n), y_n) = \sum_{k=1}^{K}\frac{N_k}{N}\underbrace{\left(\frac{1}{N_k}\sum_{n=1}^{N_k}\max_{\tilde{x}_n \in \mathcal{B}_\epsilon[x_n]} \ell(f_\theta(\tilde{x}_n^k), y_n^k)\right)}_{\mathcal{L}_k}$$

$$\geq (1+\alpha)\sum_{k=1}^{\widehat{K}}\frac{N_{\phi(k)}}{N}\mathcal{L}_{\phi(k)} + (1-\alpha)\sum_{k=\widehat{K}+1}^{K}\frac{N_{\phi(k)}}{N}\mathcal{L}_{\phi(k)}$$

$$\doteq \mathcal{L}^\alpha(\widehat{K}), \qquad \text{s.t. } \alpha \in [0,\ 1),\ \widehat{K} \leq \frac{K}{2},$$

SFAT

α-Slack Mechanism

$$\mathcal{L}_{AT} = \frac{1}{N}\sum_{n=1}^{N}\max_{\tilde{x}_n \in \mathcal{B}_\epsilon[x_n]} \ell(f_\theta(\tilde{x}_n), y_n) = \sum_{k=1}^{K}\frac{N_k}{N}\underbrace{\left(\frac{1}{N_k}\sum_{n=1}^{N_k}\max_{\tilde{x}_n \in \mathcal{B}_\epsilon[x_n]} \ell(f_\theta(\tilde{x}_n^k), y_n^k)\right)}_{\mathcal{L}_k}$$

$$\boxed{\geq (1+\alpha)\sum_{k=1}^{\widehat{K}}\frac{N_{\phi(k)}}{N}\mathcal{L}_{\phi(k)} + (1-\alpha)\sum_{k=\widehat{K}+1}^{K}\frac{N_{\phi(k)}}{N}\mathcal{L}_{\phi(k)}}$$

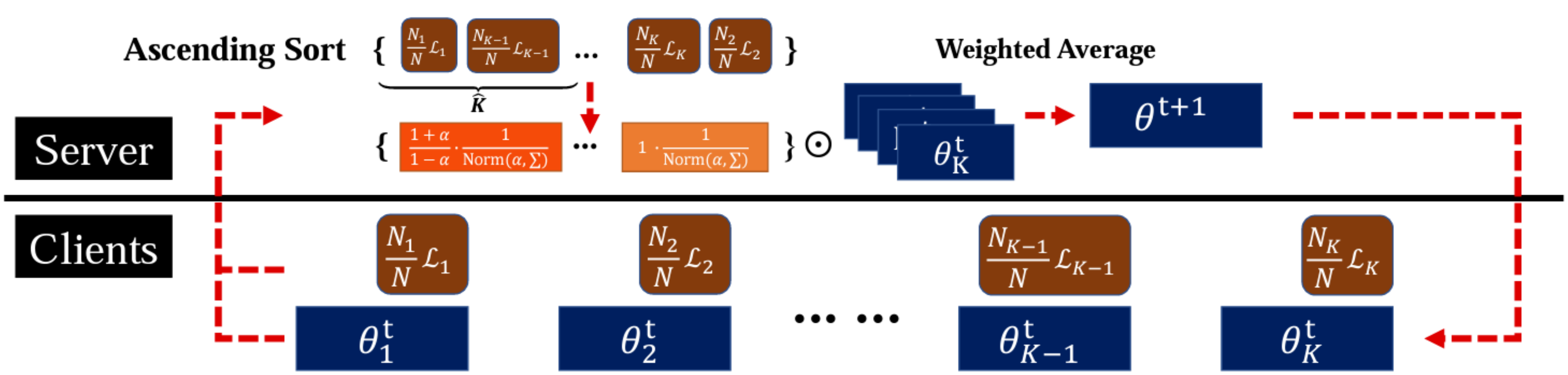
$$\doteq \mathcal{L}^\alpha(\widehat{K}), \qquad \text{s.t. } \alpha \in [0,\ 1),\ \widehat{K} \leq \frac{K}{2},$$

Theorem 4.1. *$\mathcal{L}^\alpha(\widehat{K})$ is monotonically decreasing w.r.t. both α and $\widehat{K}$, i.e., $\mathcal{L}^{\alpha_1}(\widehat{K}) < \mathcal{L}^{\alpha_2}(\widehat{K})$ if $\alpha_1 > \alpha_2$ and $\mathcal{L}^\alpha(\widehat{K}_1) < \mathcal{L}^\alpha(\widehat{K}_2)$ if $\widehat{K}_1 > \widehat{K}_2$. Specifically, $\mathcal{L}^\alpha(\widehat{K})$ recovers $\mathcal{L}$ of adversarial training when α achieves 0, and $\mathcal{L}^\alpha(\widehat{K})$ relaxes $\mathcal{L}$ to a lower bound objective by increasing $\widehat{K}$ and α.*

Slacked Federated Adversarial Training

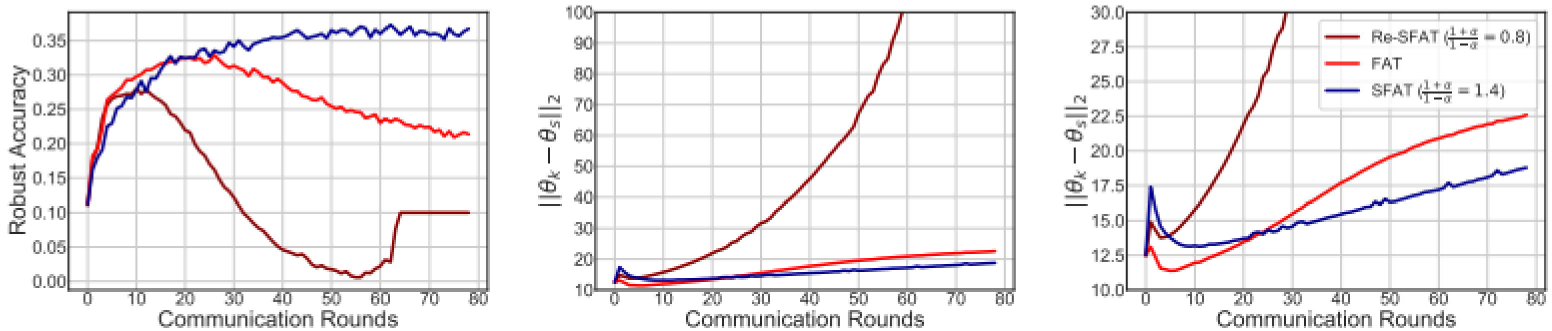
SFAT

$$\min \mathcal{L}_{\text{SFAT}} = \min_{f_\theta \in \mathcal{F}} \frac{1}{\sum_k^K N_k} \sum_{k=1}^{K} P_k N_k \cdot \underbrace{\left(\frac{1}{N_k} \sum_{n=1}^{N_k} \max_{\tilde{x}_n^k \in \mathcal{B}_\epsilon[x_n^k]} \ell(f_\theta(\tilde{x}_n^k), y_n^k) \right)}_{\mathcal{L}_k}$$

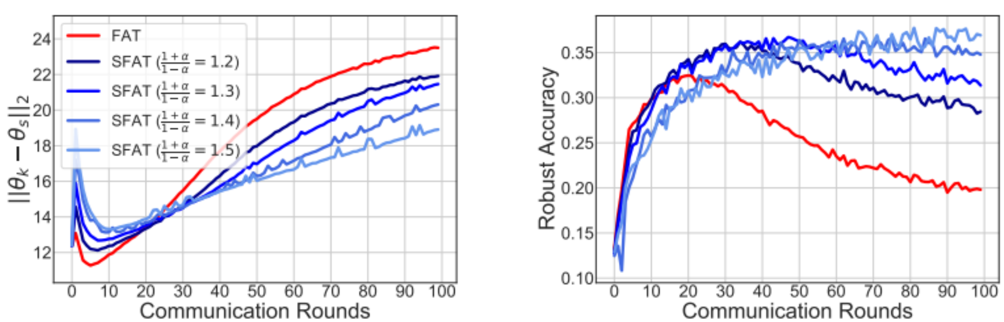


SFAT

Empirical Properties



(a) Robust accuracy (b) Client drift (c) Zoom-in Client drift

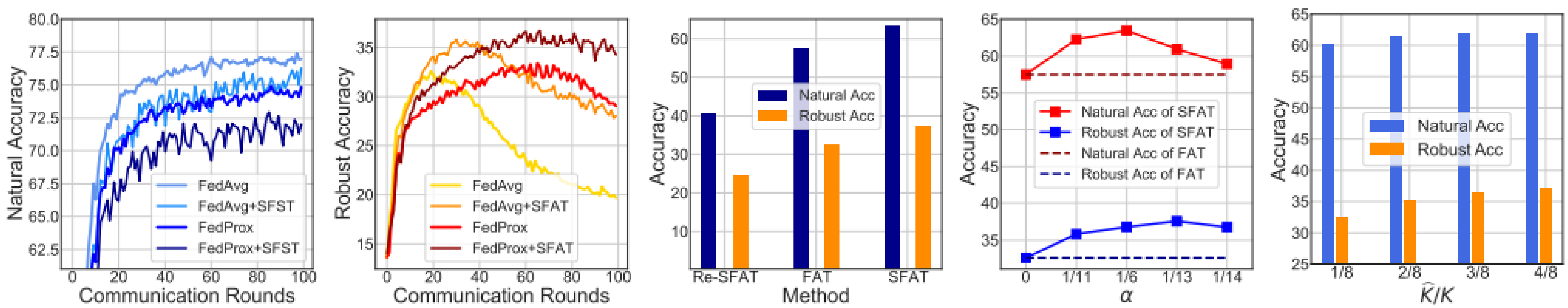


(a) Client drift (b) Robust accuracy

a. Compare SFAT with Re-SFAT

b. Present the corresponding client drift with robust acc.

Experiment



a. Ablation Study

b. Performance on different views

Table 1: Test accuracy on *CIFAR-10* (Non-IID) partition with different client numbers.

Client Number	Methods	Natural	PGD-20	CW_∞
10	FAT	56.62%	31.24%	29.82%
	SFAT	**56.67%**	**33.31%**	**31.58%**
20	FAT	60.55%	32.67%	31.07%
	SFAT	**62.24%**	**35.66%**	**33.21%**
25	FAT	58.97%	32.98%	31.14%
	SFAT	**62.73%**	**35.75%**	**33.16%**
50	FAT	56.74%	32.91%	30.50%
	SFAT	**57.21%**	**34.35%**	**31.75%**

Table 2: Test accuracy on *CIFAR-10* (Non-IID) with different local adversarial training methods.

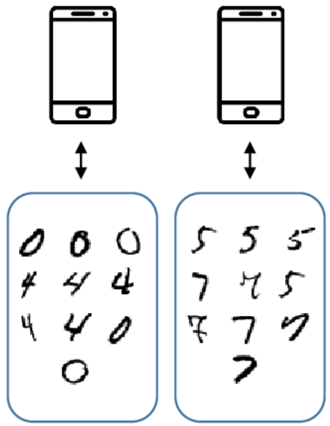
Methods		Natural	PGD-20	CW_∞
AT	FAT	57.45%	32.58%	30.52%
	SFAT	**62.34%**	**35.59%**	**33.06%**
TRADES	FAT	64.00%	31.64%	28.95%
	SFAT	**65.26%**	**35.10%**	**31.80%**
MART	FAT	56.29%	36.27%	32.41%
	SFAT	**58.41%**	**38.90%**	**34.67%**

Experiment

Table 3: Performance on Non-IID settings with different federated optimization methods (Mean±Std).

Setting		Non-IID				
CIFAR-10		Natural	FGSM	PGD-20	CW_∞	AA
FedAvg	FAT	58.13±0.68%	40.06±0.62%	32.56±0.01%	30.88±0.37%	29.17±0.03%
	SFAT	**63.36±0.07%**	**44.82±0.32%**	**37.14±0.03%**	**33.39±0.61%**	**31.66±0.70%**
FedProx	FAT	59.95±0.45%	41.44±0.15%	33.83±0.01%	31.65±0.36%	30.11±0.09%
	SFAT	**62.04±0.47%**	**44.21±0.08%**	**36.64±0.11%**	**32.62±0.20%**	**31.83±0.15%**
Scaffold	FAT	61.44±1.37%	42.85±0.76%	34.08±0.05%	32.56±0.02%	31.03±0.08%
	SFAT	**63.16±0.96%**	**45.55±0.50%**	**37.33±0.02%**	**34.82±0.04%**	**33.32±0.01%**
CIFAR-100		Natural	FGSM	PGD-20	CW_∞	AA
FedAvg	FAT	34.63±0.56%	19.92±0.28%	15.40±0.20%	13.23±0.03%	12.23±0.01%
	SFAT	**35.65±0.54%**	**20.23±0.44%**	**16.24±0.16%**	**13.53±0.02%**	**12.45±0.03%**
FedProx	FAT	31.93±0.43%	19.06±0.17%	15.30±0.08%	12.93±0.02%	12.01±0.04%
	SFAT	**34.87±0.24%**	**20.54±0.08%**	**16.09±0.10%**	**13.35±0.12%**	**12.44±0.20%**
Scaffold	FAT	39.98±0.02%	24.30±0.04%	19.34±0.07%	16.49±0.12%	15.29±0.08%
	SFAT	**44.13±0.05%**	**25.32±0.94%**	**20.22±0.07%**	**16.96±0.17%**	**15.80±0.10%**
SVHN		Natural	FGSM	PGD-20	CW_∞	AA
FedAvg	FAT	**91.52±0.28%**	88.13±0.18%	68.98±0.11%	68.04±0.15%	66.59±0.04%
	SFAT	91.26±0.01%	**88.27±0.02%**	**72.04±0.32%**	**69.96±0.16%**	**68.89±0.27%**
FedProx	FAT	91.00±0.08%	87.65±0.15%	68.48±0.04%	67.16±0.02%	65.76±0.18%
	SFAT	**91.19±0.06%**	**88.15±0.01%**	**71.84±0.30%**	**69.88±0.35%**	**68.84±0.37%**
Scaffold	FAT	90.82±0.87%	87.89±0.66%	69.51±0.84%	68.12±0.88%	67.19±0.54%
	SFAT	**90.93±0.76%**	**88.27±0.45%**	**71.77±0.38%**	**69.49±0.67%**	**68.37±0.48%**

Non-IID dataset



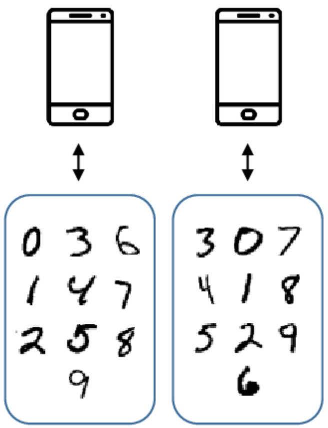
In the performance comparison, we demonstrate the effectiveness of SFAT in Three datasets.

Experiment

Table 21: Performance on three benchmark datasets under different federated optimization methods (Non-IID & IID).

Setting		Non-IID					IID				
CIFAR-10		Natural	FGSM	PGD-20	CW_∞	AA	Natural	FGSM	PGD-20	CW_∞	AA
Centralized AT		-	-	-	-	-	66.47%	47.68%	38.18%	37.04%	34.48%
FedAvg	FAT	57.45%	39.44%	32.58%	30.52%	29.20%	**69.35%**	48.45%	37.43%	35.72%	33.96%
	SFAT	**63.44%**	**45.13%**	**37.17%**	**33.99%**	**32.36%**	67.43%	**50.33%**	**42.78%**	**37.91%**	**36.20%**
FedProx	FAT	60.44%	41.59%	33.84%	31.29%	30.02%	66.91%	46.70%	37.14%	34.54%	32.68%
	SFAT	**62.51%**	**44.29%**	**36.75%**	**33.82%**	**31,98%**	**68.31%**	**48.40%**	**42.41%**	**37.25%**	**35.97%**
Scaffold	FAT	62.81%	43.61%	34.13%	32.53%	30.95%	68.27%	49.25%	39.33%	37.31%	35.30%
	SFAT	**64.12%**	**46.05%**	**37.35%**	**34.78%**	**33.32%**	**71.36%**	**50.42%**	**43.83%**	**39.12%**	**35.47%**
CIFAR-100		Natural	FGSM	PGD-20	CW_∞	AA	Natural	FGSM	PGD-20	CW_∞	AA
Centralized AT		-	-	-	-	-	35.81%	23.09%	18.64%	16.48%	15.42%
FedAvg	FAT	35.19%	20.20%	15.60%	13.26%	12.22%	32.65%	20.44%	16.47%	14.10%	12.99%
	SFAT	**36.18%**	**20.70%**	**16.40%**	**13.55%**	**12.42%**	**38.36%**	**21.86%**	**17.10%**	**14.36%**	**13.42%**
FedProx	FAT	32.36%	19.22%	15.37%	12.91%	12.05%	34.78%	20.71%	16.37%	14.28%	13.09%
	SFAT	**35.11%**	**20.62%**	**16.19%**	**13.47%**	**12.63%**	**37.58%**	**21.74%**	**17.03%**	**14.46%**	**13.50%**
Scaffold	FAT	39.96%	24.26%	19.41%	16.60%	15.37%	43.80%	26.25%	20.76%	18.39%	17.20%
	SFAT	**44.08%**	**24.38%**	**20.29%**	**16.79%**	**15.90%**	**44.36%**	**28.65%**	**23.14%**	**20.11%**	**18.39%**
SVHN		Natural	FGSM	PGD-20	CW_∞	AA	Natural	FGSM	PGD-20	CW_∞	AA
Centralized AT		-	-	-	-	-	92.39%	89.75%	72.73%	72.31%	70.93%
FedAvg	FAT	91.24%	87.95%	68.87%	67.89%	66.54%	**93.52%**	**90.68%**	72.24%	71.22%	70.08%
	SFAT	**91.25%**	**88.28%**	**71.72%**	**69.79%**	**68.62%**	92.75%	90.06%	**74.37%**	**72.34%**	**71.27%**
FedProx	FAT	90.92%	87.50%	68.44%	67.18%	65.94%	93.54%	90.66%	72.53%	71.42%	70.21%
	SFAT	**91.25%**	**88.15%**	**71.54%**	**69.53%**	**68.47%**	**93.59%**	**90.80%**	**74.66%**	**72.67%**	**71.48%**
Scaffold	FAT	89.95%	87.23%	68.66%	67.23%	66.65%	93.80%	91.00%	73.26%	72.05%	70.80%
	SFAT	**90.20%**	**87.81%**	**71.39%**	**68.81%**	**67.88%**	**93.92%**	**91.28%**	**75.96%**	**74.05%**	**72.88%**

IID dataset

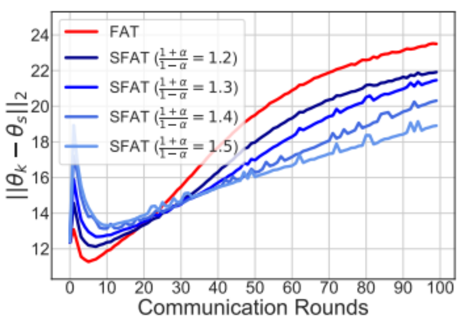


We find that our method could improve the original FAT on not only Non-IID but also IID setting in FL.

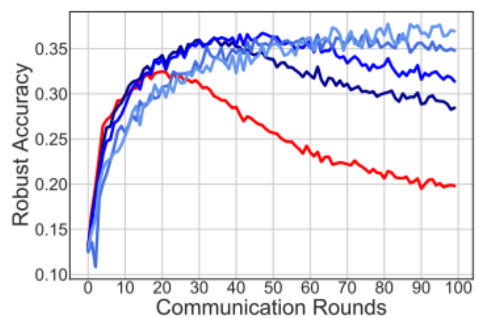
Future: Intensified Heterogeneity

Does Intensified Heterogeneity in aggregating robust models has been already fixed?

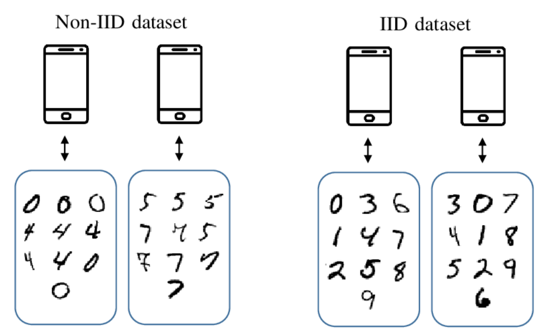
The answer is: <u>Not Yet!</u>



(a) Client drift



(b) Robust accuracy



Q1: How to solve such algorithmic conflict?

Q2: Why it also happens on IID dataset?

Summary

- In this work, We study the critical **robustness deterioration** in FAT, and discover that the reason behind this phenomenon may attribute to the **intensified data heterogeneity** induced by the adversarial generation in local clients.

- We derive an **α-slack mechanism** for adversarial training to relax the inner-maximization to a lower bound for SFAT, which could asymptotically approach the original goal towards adversarial robustness and alleviate the intensified heterogeneity in federated learning.

Thank You!

OpenReview

Code