FedCART: Tackling Long-Tailed Distributions in Federated Adversarial Training via Classifier Refinement

Yuchen Qin[1], Yizhi Zhou[2], Junxiao Wang[3], Xin Xie[4], Heng Qi[1*]

[1]Dalian University of Technology, [2]Dalian Ocean University,

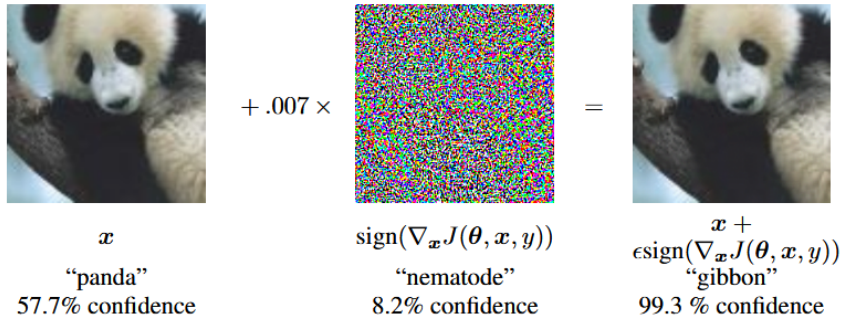
[3]Guangzhou University, [4]Tianjin University

2018qyc@mail.dlut.edu.cn, hengqi@dlut.edu.cn

Codes: github.com

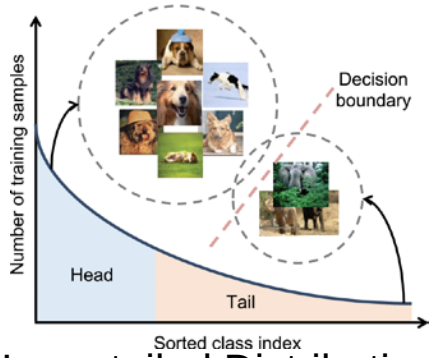
Realistic Concern

- Federated Learning is vulnerable to adversarial perturbations.
- Non-IID data across client makes training challenging.
- Open-world scenarios often feature data with long-tailed distributions.



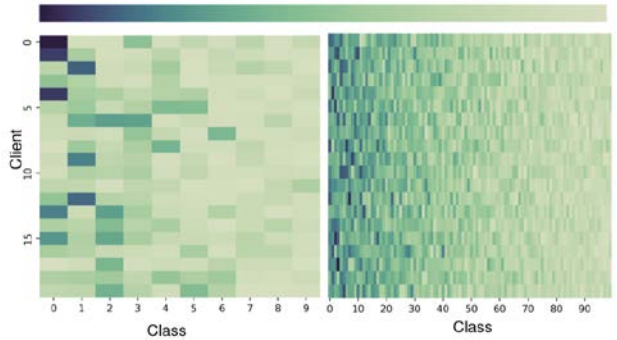
Adversarial Perturbations

(Goodfellow et al. 2014)



Long-tailed Distribution

(Zhang et al. 2023)

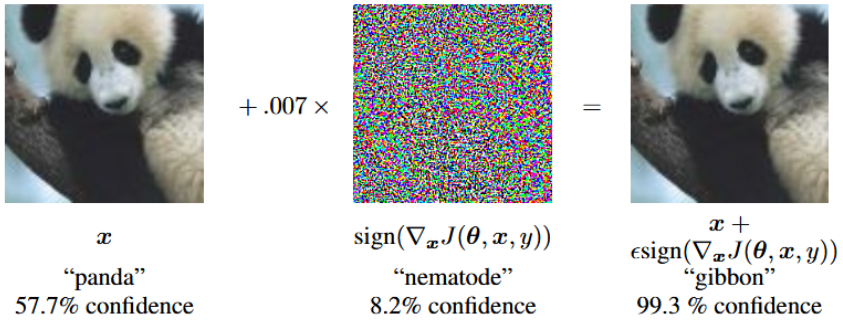


Non-IID Data across Clients

(Xiao et al. 2023)

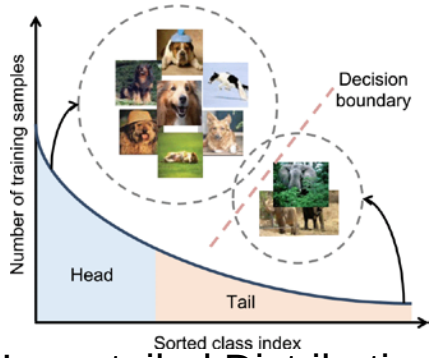
When FAT meets Long-tailed data

- **Problem:** Existing Federated Adversarial Training(FAT) assume a global class-balanced data, a condition that rarely holds in the real world. FAT under long-tailed distribution remains largely unexplored.
- **Input:** Global long-tailed and Non-IID training data across clients & balanced test data.
- **Goal:** To make a shared adversarial robust model can generalize well.



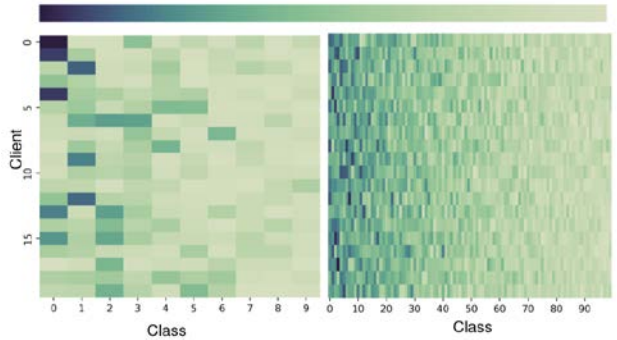
Adversarial Perturbations

(Goodfellow et al. 2014)



Long-tailed Distribution

(Zhang et al. 2023)

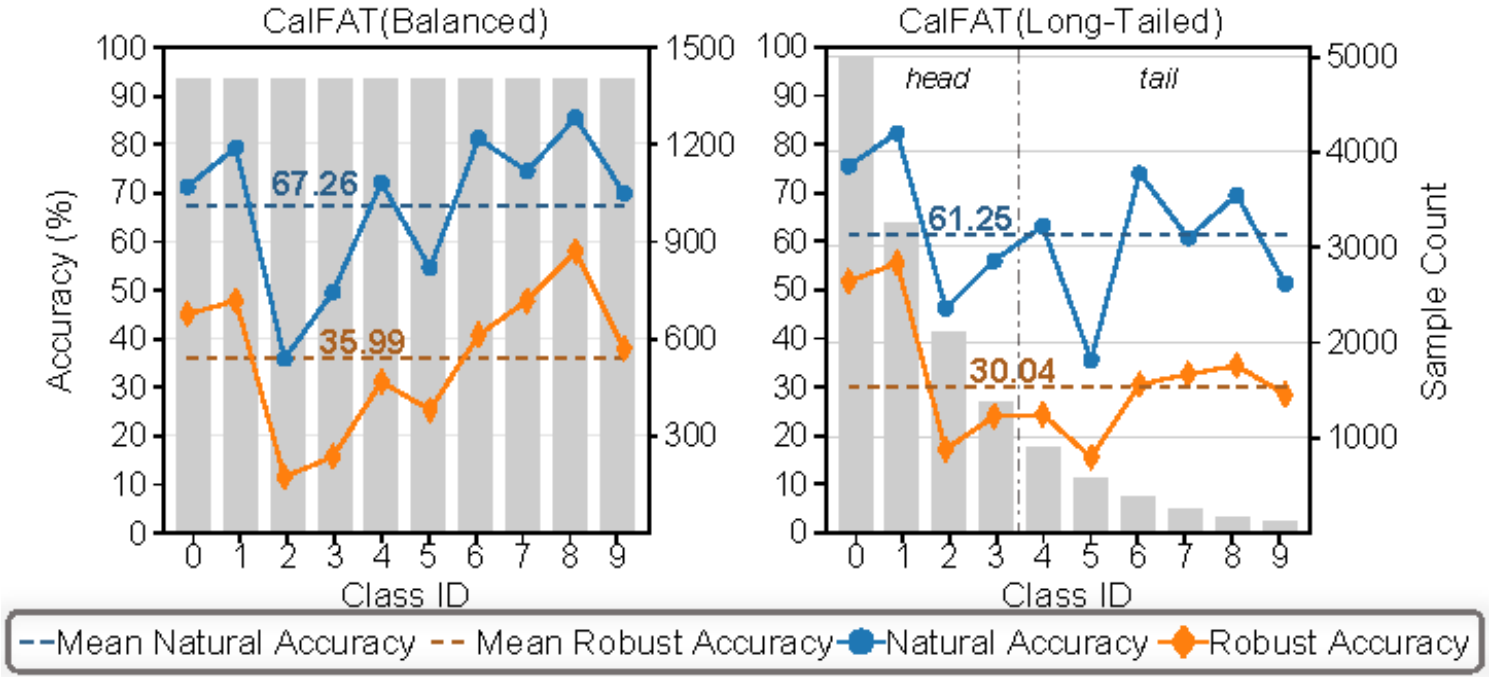


Non-IID Data across Clients

(Xiao et al. 2023)

Unexpected Performance Degraded.

- Calibrated Federated Adversarial Training (CalFAT, Zhang et al. 2023) tackle the Non-IID by calibrating the logits adaptively under global balanced data.
- Long-tailed data bias the prediction to the head classes and lead performance degraded.

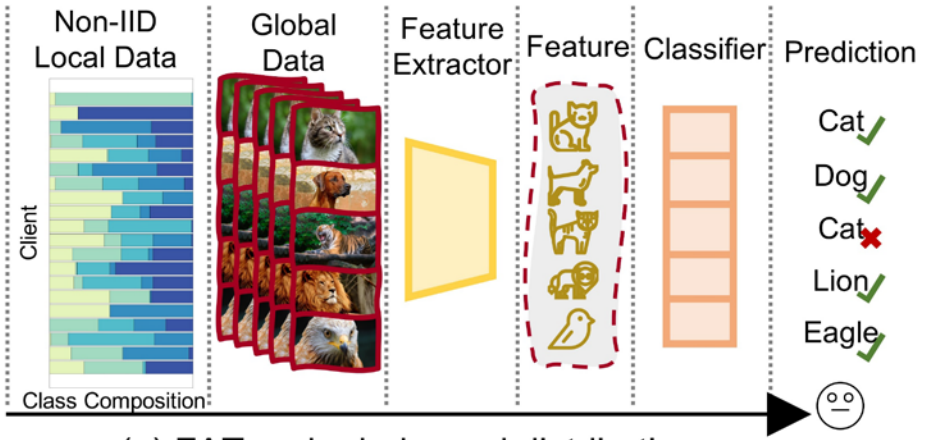


Performance of CalFAT under **Left:** class-balanced and **Right:** long-tailed distribution.

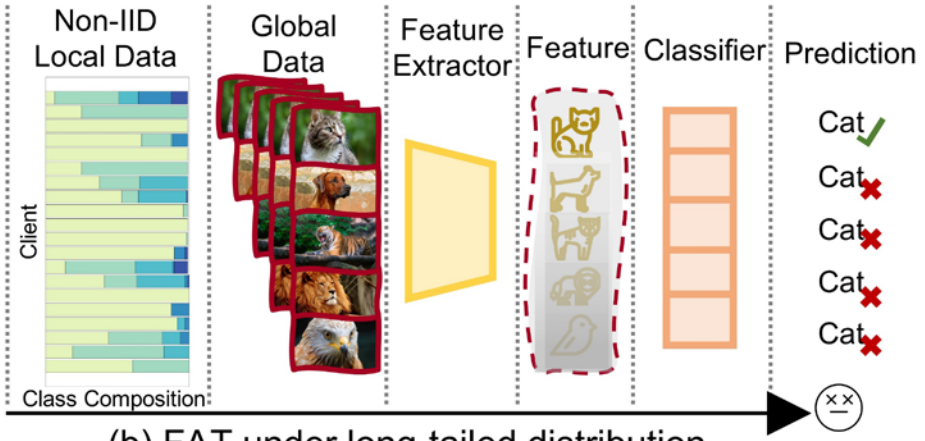
When FAT meets Long-tailed data

Why does performance degrade?

- **Flawed Assumption.** Balanced-data assumption rarely holds in practice, but long-tailed data in the real-world is often ignored.
- **Underrepresented Features**. Tail classes concentrate on limited clients, intensifying inter-client inconsistency. This renders tail-class features highly ambiguous.
- **Confused Classifier.** The extreme scarcity of tail-class samples biases local models and skews feature representations, ultimately confusing the classifier.



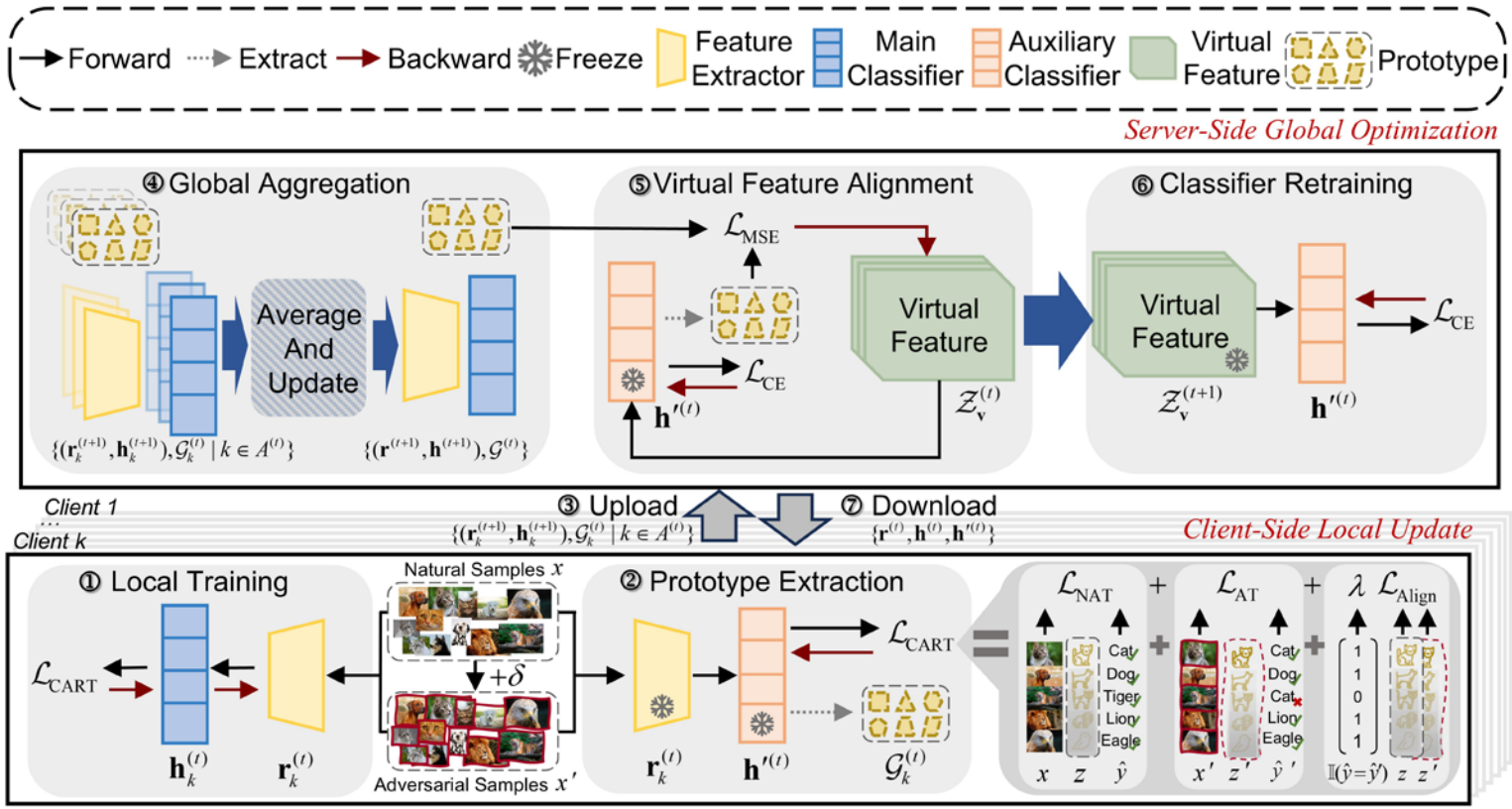
(a) FAT under balanced distribution



(b) FAT under long-tailed distribution

The illustration of FAT under **(a)** balanced and **(b)** long-tailed data.

How to tackle the challenges? FedCART.



The framework of FedCART, a simple yet effective framework with a shared feature extractor and a dual-classifier.

Local Robust Update

- **Contrastive Alignment:** Local updates guided by $\mathcal{L}_{\text{CART}}$, a feature-alignment loss designed to fortify models against adversarial drifts.

$$\mathcal{L}_{\text{CART}} = \mathcal{L}_{\text{NAT}} + \mathcal{L}_{\text{AT}} + \gamma \mathcal{L}_{\text{Align}}$$

- **Prototype Extraction:** Each client computes the set of Gradient-based Class Prototypes $\mathcal{G}_k^{(t)} = \{\mathrm{g}_{k,c}^{(t)} \mid c \in C_k\}$ to capture local update signatures.

$$\mathrm{g}_{k,c}^{(t)} = \frac{1}{n_k^{(c)}} \sum_{i=1}^{n_k^{(c)}} \nabla_{\mathrm{h}'} \mathcal{L}_{\text{CART}} \left(\mathrm{w}_{\mathrm{k}}^{\prime(\mathrm{t})}; \mathrm{x}_{\mathrm{i}}, \mathrm{x}_{\mathrm{i}}^{\prime}, \mathrm{y}_{\mathrm{i}} \right).$$

Algorithm 1 The Update Procedure of Client k

Input: Global model parameters $\mathbf{w}^{(t)} = (\mathbf{r}^{(t)}, \mathbf{h}^{(t)},)$, and auxiliary classifier parameters $\mathbf{h}'^{(t)}$.

Output: Updated parameters $\mathbf{w}_k^{(t+1)} = (\mathbf{r}_k^{(t+1)}, \mathbf{h}_k^{(t+1)})$ and the set of gradient-based prototypes $\mathcal{G}_k^{(t)}$.

Procedure: Client$(\mathbf{r}^{(t)}, \mathbf{h}^{(t)}, \mathbf{h}'^{(t)})$:

Initialize local model $\mathbf{w}_k^{(t)} = (\mathbf{r}^{(t)}, \mathbf{h}^{(t)})$.
// Local training
for local epoch 1 to E **do**
  Update local model $\mathbf{w}_k^{(t)}$ following $\mathcal{L}_{\text{CART}}$.
end for
// Extraction of class prototypes
Initialize local model $\mathbf{w}'^{(t)}_k = (\mathbf{r}^{(t)}, \mathbf{h}'^{(t)})$.
for each class $c = 1$ to C_k **do**
  Extract $g_{k,c}^{(t)}$ using Eq. (6).
end for
return $\mathbf{w}_k^{(t+1)} = (\mathbf{r}_k^{(t+1)}, \mathbf{h}_k^{(t+1)})$ and $\mathcal{G}_k^{(t)}$.

Global Model Refinement

- **Global Aggregation:** Server performs aggregation for both models and gradient-based prototypes.

$$\bar{\mathrm{g}}_c^{(t)} = \sum_{k \in A^{(t)}} \frac{N_k}{\sum_{k' \in A^{(t)}} N_{k'}} \mathrm{g}_{k,c}^{(t)}.$$

- **Virtual Feature Alignment:** Server update virtual feature so that their gradient are aligned with the prototypes.

$$\mathcal{L}_{\mathrm{MSE}}(\mathcal{Z}_{\mathrm{v}}^{(t)}; \mathrm{h}'^{(t)}, \mathcal{G}^{(t)}) = \left\| \nabla_{\mathbf{h}'} \mathcal{L}_{\mathrm{CE}}(\mathrm{h}'^{(t)}; \mathcal{Z}_{\mathrm{v}}^{(t)}) - \mathcal{G}^{(t)} \right\|^2.$$

$$\mathcal{Z}_{\mathrm{v}}^{(t+1)} \leftarrow \mathcal{Z}_{\mathrm{v}}^{(t)} - \hat{\eta}_{a'} \nabla_{\mathrm{z}} \mathcal{L}_{\mathrm{MSE}}(\mathcal{Z}_{\mathrm{v}}^{(t)}; \mathrm{h}'^{(t)}, \mathcal{G}^{(t)}).$$

- **Bias Calibration:** Refining the auxiliary classifier on balanced virtual features to calibrate the skewness.

$$\mathrm{h}'^{(t+1)} \leftarrow \mathrm{h}'^{(t)} - \hat{\eta}_{r'} \nabla_{\mathbf{h}'} \mathcal{L}_{\mathrm{CE}}(\mathrm{h}'^{(t)}; \mathcal{Z}_{\mathrm{v}}^{(t)}).$$

Algorithm 2 Training Procedure of FedCART

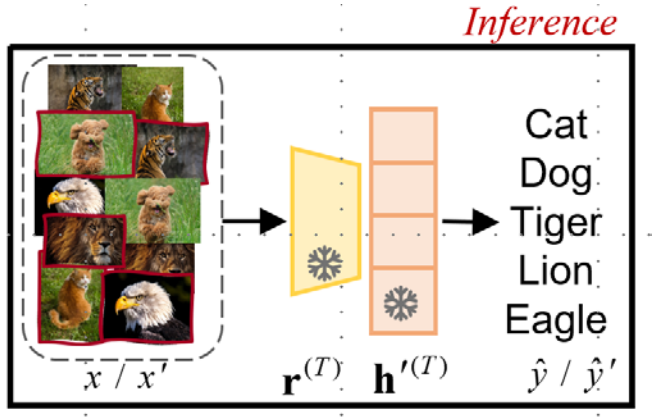
Input: Initialized global model parameters $\mathbf{w}^{(0)} = (\mathbf{r}^{(0)}, \mathbf{h}^{(0)})$, auxiliary classifier parameters $\mathbf{h}'^{(0)}$, number of total communication rounds T, number of virtual feature update epochs T_V, and number of auxiliary classifier re-training epochs T_R.

Output: Trained global model $\mathbf{w}^{(T)} = (\mathbf{r}^{(T)}, \mathbf{h}'^{(T)})$.

for $t = 1$ **to** T **do**
  Randomly select active client set $A^{(t)}$.
  `// Client-side local update`
  for all each client $k \in A^{(t)}$ (in parallel) **do**
    $\{\mathbf{w}_k^{(t+1)}, \mathcal{G}_k^{(t)}\} \leftarrow \mathrm{Client}_k(\mathbf{r}^{(t)}, \mathbf{h}^{(t)}, \mathbf{h}'^{(t)})$.
  end for
  `// Server-side global optimization`
  Aggregate $\mathbf{w}^{(t+1)}$ and $\bar{\mathbf{g}}_c^{(t)}$ using Eq. (1) and Eq. (7).
  Update virtual feature set $\mathcal{Z}_{\mathrm{v}}^{(t)}$ using Eq. (8) and Eq. (9) for T_V epochs.
  Retrain auxiliary classifier $\mathbf{h}'^{(t)}$ using Eq. (10) for T_R epochs.
  Broadcast $(\mathbf{r}^{(t+1)}, \mathbf{h}^{(t+1)}, \mathbf{h}'^{(t+1)})$ to active clients.
end for

Why does FedCART work?

- **Central Idea:** FedCART aligns natural and adversarial feature representations on clients and refine the classifier via a set of balanced virtual features on the server, thereby mitigating the performance degradation induced by long-tailed, heterogeneous data in FAT.
- **Inference:** Only auxiliary classifier with the final extractor for inference.



The illustration of FedCART's inference.

Settings

- **Dataset**(-LT)： CIFAR10, CIFAR100, FMNIST, SVHN.
- **Partition**:
  - Global long-tailed (Cao et al. 2019) .
  - Non-IID (Chen et al. 2022) across clients.
- **Metrics**:
  - Natural accuracy (Natural).
  - Robust accuracy under adversarial attacks (FGSM, PGD, CW, AA) and their average (Robust AVG).
- **Baselines**:
  - FedFAT
  - CalFAT

  } FAT methods.

  - FedPGD
  - FedTRADES
  - FedMART
  - FedTAET

  } FedAvg + centralized adversarial training (CAT).

Main Result

- FedCART achieves the superior performance across all datasets.

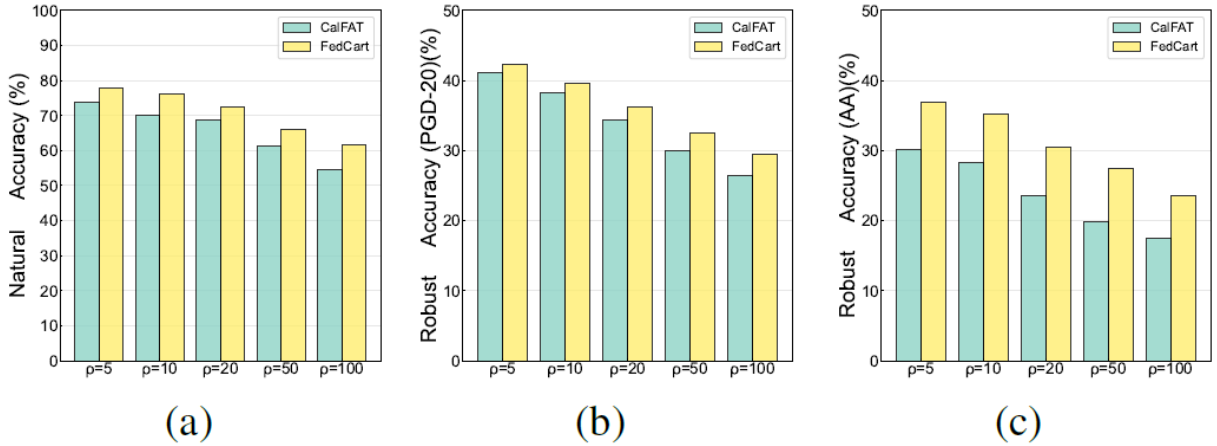
Dataset	CIFAR10-LT						CIFAR100-LT					
Metric	Natural	Robust Accuracy				Robust AVG	Natural	Robust Accuracy				Robust AVG
		FGSM [6]	CW [2]	PGD-20 [20]	AA [5]			FGSM [6]	CW [2]	PGD-20 [20]	AA [5]	
FedPGD [20]	31.24	22.76	21.03	21.65	20.65	21.52	26.65	13.40	11.08	12.20	10.50	11.80
FedMART [34]	26.83	22.07	20.72	21.61	20.46	21.21	20.30	12.64	10.68	12.22	10.34	11.47
FedTAET [40]	33.98	23.62	21.49	22.42	20.80	22.08	30.41	13.22	9.84	11.85	8.91	10.96
FedTRADES [41]	41.59	26.30	23.74	25.02	23.49	24.64	24.85	14.72	12.29	14.04	11.94	13.25
FedFAT [50]	40.16	26.47	24.15	24.69	23.54	24.71	31.78	14.55	11.93	12.67	11.13	12.57
CalFAT [3]	61.25	32.86	21.08	30.04	19.93	25.98	36.86	15.85	9.18	14.00	8.40	11.86
FedCART (ours)	**66.12**	**36.59**	**30.29**	**32.47**	**27.47**	**31.71**	**38.19**	**17.37**	**14.04**	**15.65**	**12.53**	**14.90**

Dataset	SVHN-LT						FMNIST-LT					
Metric	Natural	Robust Accuracy				Robust AVG	Natural	Robust Accuracy				Robust AVG
		FGSM [6]	CW [2]	PGD-20 [20]	AA [5]			FGSM [6]	CW [2]	PGD-20 [20]	AA [5]	
FedPGD [20]	84.23	58.90	52.60	51.81	51.31	53.66	76.52	68.47	67.53	67.70	67.32	67.75
FedMART [34]	84.78	63.31	57.88	57.77	56.94	58.98	75.58	67.33	66.28	66.55	66.07	66.56
FedTAET [40]	78.30	62.70	59.07	60.98	58.52	60.32	80.21	69.34	67.66	68.16	67.30	68.11
FedTRADES [41]	77.67	63.18	59.69	60.59	59.05	60.63	80.71	68.49	66.35	66.44	65.72	66.75
FedFAT [50]	84.86	60.16	53.78	53.03	52.53	54.88	80.85	69.59	68.07	68.25	67.74	68.41
CalFAT [3]	85.93	66.59	55.63	62.86	54.72	59.95	**86.75**	73.61	65.89	71.72	65.57	69.20
FedCART (ours)	**90.98**	**74.99**	**69.73**	**69.29**	**68.06**	**70.52**	85.13	**74.75**	**72.95**	**73.64**	**72.68**	**73.51**

Natural and robust accuracies(%) across different datasets..

Effect of Long-tailed

- FedCART exhibits outstanding performance with various degree of long-tailed, with the significant improvement from the Medium and Minority class.



The (a) Natural accuracy, (b) Robust accuracy under PGD-20, and (c) Robust accuracy under AA of CalFAT and FedCART under different degree of long-tailed on CIFAR10-LT.

Method	Majority		Medium		Minority	
	Natural	PGD-20 [20]	Natural	PGD-20 [20]	Natural	PGD-20 [20]
FedPGD [20]	53.48	27.52	24.73	9.70	6.78	2.00
FedMART [34]	45.91	<u>29.80</u>	15.67	8.19	3.51	1.38
FedTAET [40]	<u>55.41</u>	25.19	30.82	10.27	10.22	2.50
FedTRADES [41]	53.08	**32.93**	21.27	10.31	5.15	1.93
FedFAT [50]	**60.45**	28.05	30.81	10.17	9.73	2.42
CalFAT [3]	50.17	20.70	<u>37.35</u>	<u>13.61</u>	<u>25.90</u>	<u>8.99</u>
FedCART (ours)	52.16	24.25	**39.51**	**15.34**	**26.06**	**9.09**

Natural and robust accuracies(%) across different classes on CIFAR100-LT.

Analysis on Data Partition

- The results in different settings of Non-IID and client participation validate the reliability of FedCART.

Non-IID	Method	Natural	PGD-20 [20]	AA [5]
$\beta = 0.25$	CalFAT [3]	**55.65**	27.35	18.00
	FedCART	51.52	**30.58**	**24.72**
$\beta = 0.5$	CalFAT [3]	61.25	30.04	19.93
	FedCART	**66.12**	**32.47**	**27.47**
$\beta = 0.75$	CalFAT [3]	63.25	30.22	20.06
	FedCART	**69.29**	**33.16**	**29.28**

Natural and robust accuracies(%) across different degree of Non-IID on CIFAR10-LT

client number		$K = 5$			$K = 20$		
active rates	Method	Natural	PGD-20 [20]	AA [5]	Natural	PGD-20 [20]	AA [5]
0.1	CalFAT [3]	–	–	–	33.71	20.13	12.89
	FedCART	–	–	–	**34.34**	**21.70**	**19.10**
0.2	CalFAT [3]	38.11	19.59	13.37	30.75	19.57	11.92
	FedCART	**44.38**	**21.46**	**18.77**	**42.95**	**25.04**	**20.85**
0.5	CalFAT [3]	**53.48**	26.83	17.46	40.04	21.04	9.92
	FedCART	51.41	**27.22**	**23.25**	**43.38**	**23.21**	**18.23**
0.7	CalFAT [3]	54.10	26.32	17.68	35.94	20.59	10.85
	FedCART	**63.00**	**27.32**	**24.10**	**46.33**	**26.05**	**21.42**
1.0	CalFAT [3]	61.30	31.30	21.05	43.32	20.99	9.98
	FedCART	**67.24**	**32.57**	**27.90**	**47.06**	**27.18**	**22.75**

Natural and robust accuracies(%) across different number of clients and active rates on CIFAR10-LT.

Combining with other methods

- The improvement, both natural and robust accuracies, from the combination of CAT and FedCART confirm the effectiveness and extensibility of our framework.

Metric	Natural	PGD-20
FedMART	26.83	21.61
MART + FedCART (ours)	33.05	27.04
FedTAET	33.98	22.42
TAET + FedCART (ours)	46.61	26.57
FedTRADES	41.59	25.02
TRADES + FedCART (ours)	65.48	27.93
FedCART (ours)	66.12	32.47

The natural and robust accuracies(%) of the combination of our framework with different CAT methods on CIFAR10-LT.

Ablation Result

- The ablation results confirm the necessity and effectiveness of each component.

Case	Configuration					Natural	PGD-20 [20]	AA [5]
	$\mathcal{L}_{NAT}$	$\mathcal{L}_{AT}$	$\mathcal{L}_{Align}$	CART	h′			
1	✓			✓	✓	**79.08**	0.00	0.00
2		✓		✓	✓	53.86	29.79	26.72
3	✓		✓	✓	✓	75.30	2.91	2.22
4		✓	✓	✓	✓	49.91	32.54	26.98
5	✓	✓		✓	✓	66.04	29.33	26.12
6	✓	✓	✓			40.95	24.58	23.48
7	✓	✓	✓	✓		49.98	28.30	25.69
Ours (Full)	✓	✓	✓	✓	✓	67.24	**32.57**	**27.90**

The accuracy(%) of ablation study for key components of FedCART on CIFAR10-LT

Conclusion

- ✓ Problem identification. We are the first to reveal and diagnose the severe failure of Federated Adversarial Training under long-tailed distributions. This is a critical gap that prior work completely overlooked.
- ✓ Novel framework. We propose FedCART, which decouples robust feature alignment on the client side from unbiased classifier retraining on the server side. The dual-classifier structure and gradient prototyping mechanism effectively eliminate long-tailed bias while preserving privacy.
- ✓ Extensive validation. Through comprehensive experiments, we show that FedCART significantly outperforms state-of-the-art methods, especially in protecting vulnerable tail classes.
- ✓ Our framework is easy to be implemented and integrated into existing works.

Conclusion

- ✓ Problem identification. We are the first to reveal and diagnose the severe failure of Federated Adversarial Training under long-tailed distributions. This is a critical gap that prior work completely overlooked.
- ✓ Novel framework. We propose FedCART, which decouples robust feature alignment on the client side from unbiased classifier retraining on the server side. The dual-classifier structure and gradient prototyping mechanism effectively eliminate long-tailed bias while preserving privacy.
- ✓ Extensive validation. Through comprehensive experiments, we show that FedCART significantly outperforms state-of-the-art methods, especially in protecting vulnerable tail classes.
- ✓ Our framework is easy to be implemented and integrated into existing works.

Thank You!

Contact: 2018qyc@mail.dlut.edu.cn, hengqi@dlut.edu.cn
Code: github.com

QR QR QR

FedCART: Tackling Long-Tailed Distributions in Federated Adversarial Training via Classifier Refinement

Yuchen Qin[1], Yizhi Zhou[2], Junxiao Wang[3], Xin Xie[4], Heng Qi[1*]

[1]Dalian University of Technology, [2]Dalian Ocean University,

[3]Guangzhou University, [4]Tianjin University

2018qyc@mail.dlut.edu.cn, hengqi@dlut.edu.cn

Codes: github.com