WeBank 微众银行
UNIVERSITI MALAYA

Ferrari: Federated Feature Unlearning via Optimizing Feature Sensitivity

Paper Code

Hanlin Gu[2*] Win Kent Ong[1*] Chee Seng Chan[1] Lixin Fan[2]
Universiti Malaya[1] WeBank AI Lab[2]

Email Linked In

NEURAL INFORMATION PROCESSING SYSTEMS

Motivation

- Previous works on federated unlearning have focused on **client**, **sample**, or **class** level unlearning.
- **Feature unlearning** in federated settings **has not been explored**.

Challenges

1. Centralized unlearning methods **impractical in federated settings**:
   - Full training datasets with participation of all clients
2. Difficulty in **evaluating the effectiveness** of feature unlearning.
   - Conventional method compared to the retrained model reduced model utility.

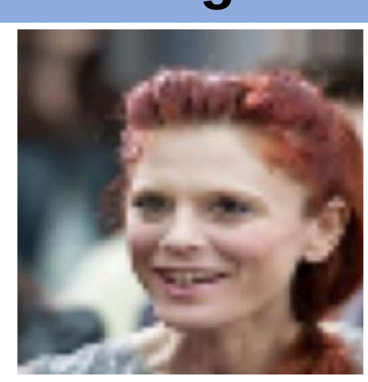
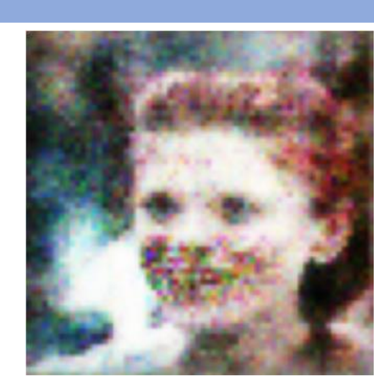
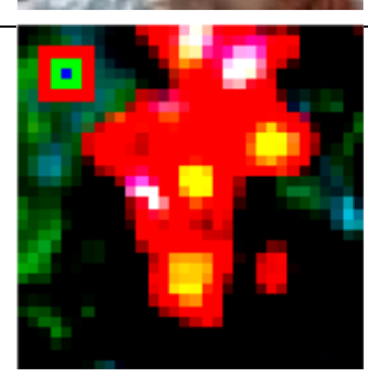
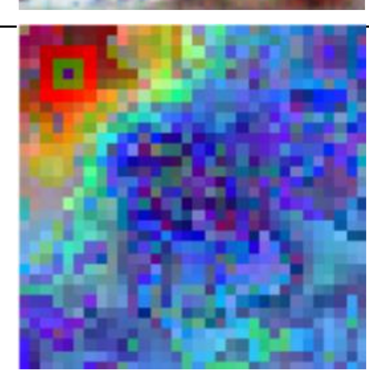
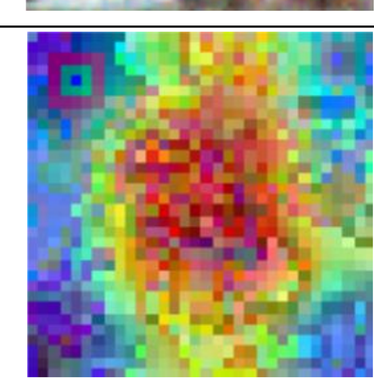
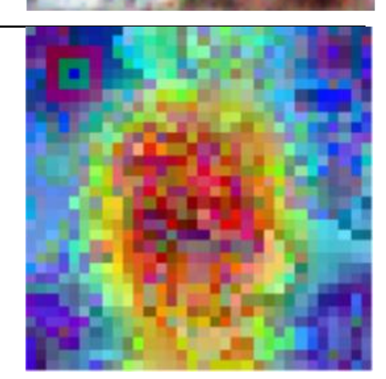
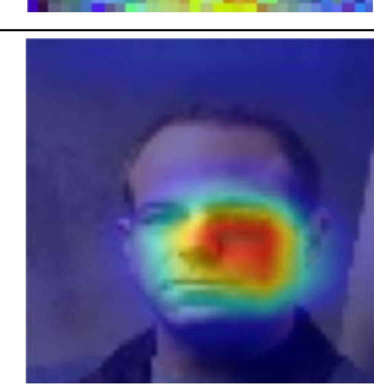
Contributions

I. We define the **Feature Sensitivity** based on Lipschitz Continuity and introduce this metric in federated feature unlearning.

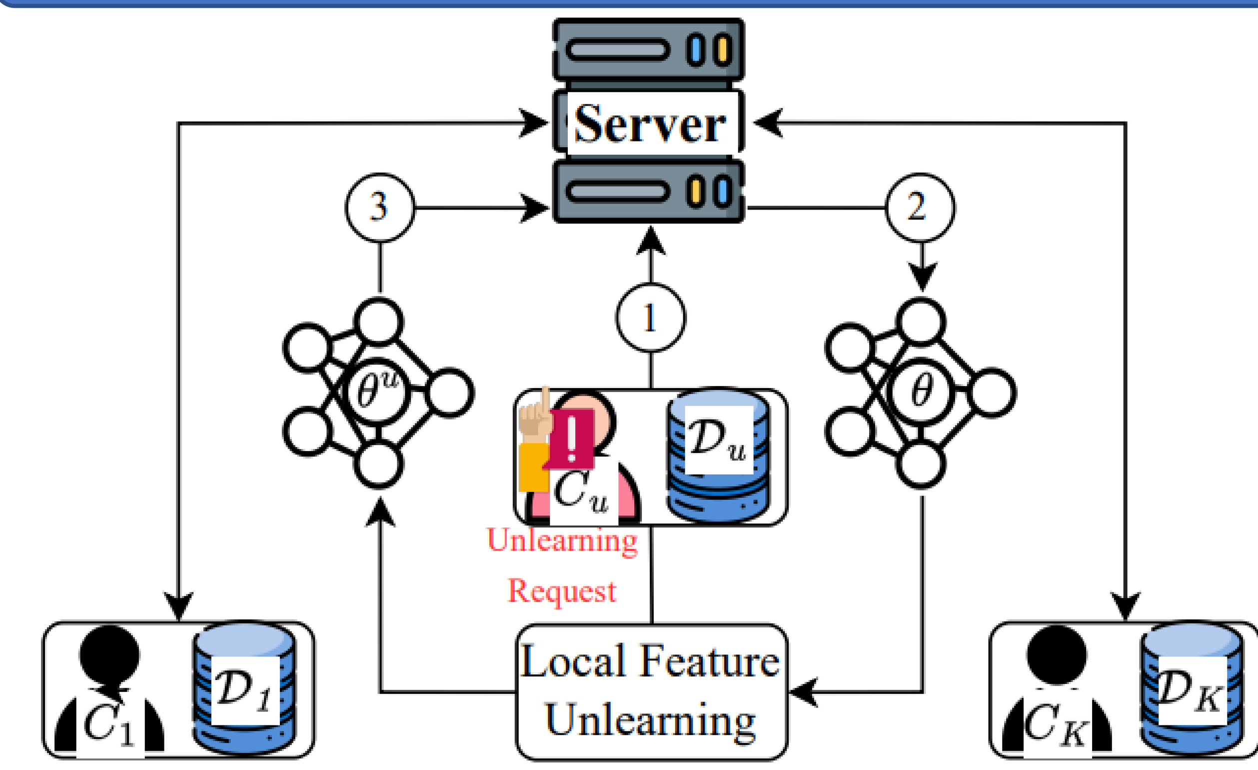
II. We proposed an effective federated feature unlearning framework, called **Ferrari**, allowing clients to **selectively unlearn specific features** from the trained global model **without the participation of other clients** by **optimizing feature sensitivity locally**.

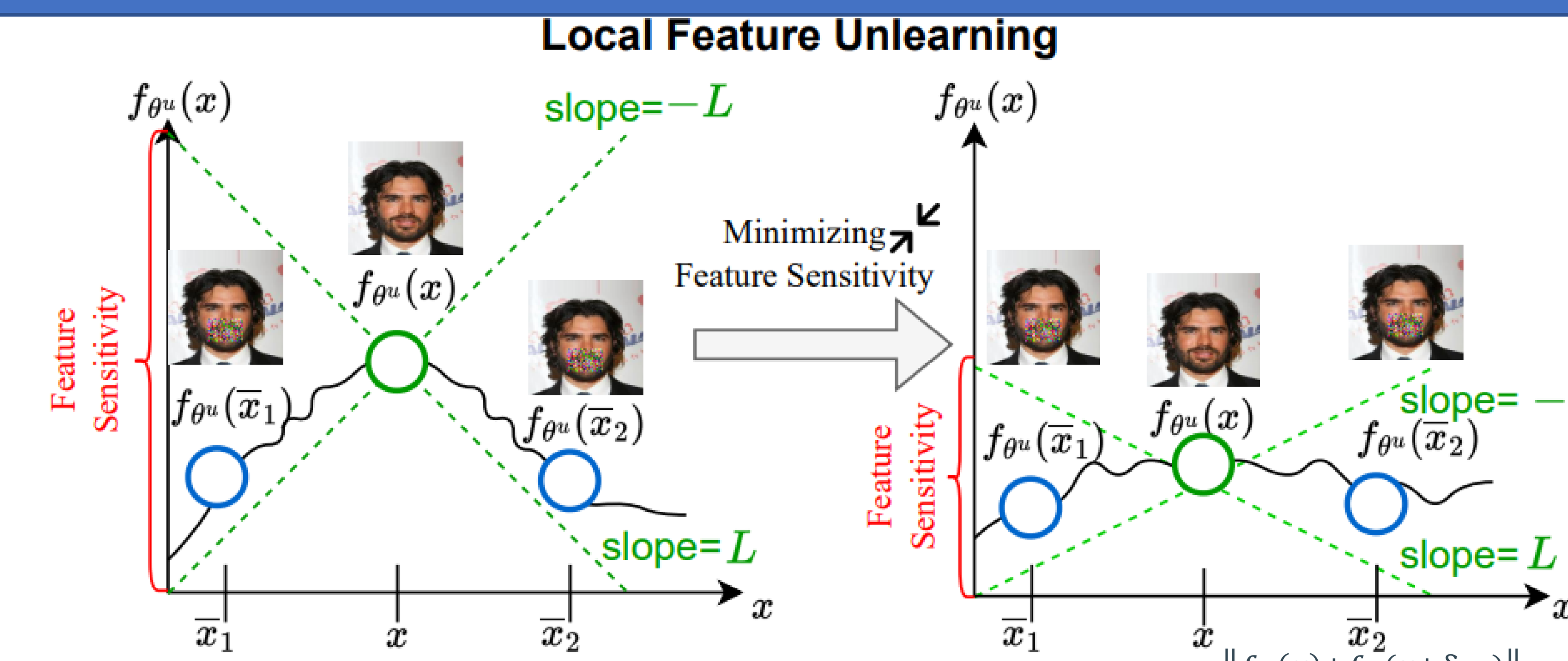
III. We provide **theoretical proof** and **experiments** showing the state-of-the-art **utility** and **effectiveness** of our framework.

Qualitative Results

Scenario	Target	Baseline	Retrain	Ours
Sensitive				
Backdoor				
Biased				

Proposed Method





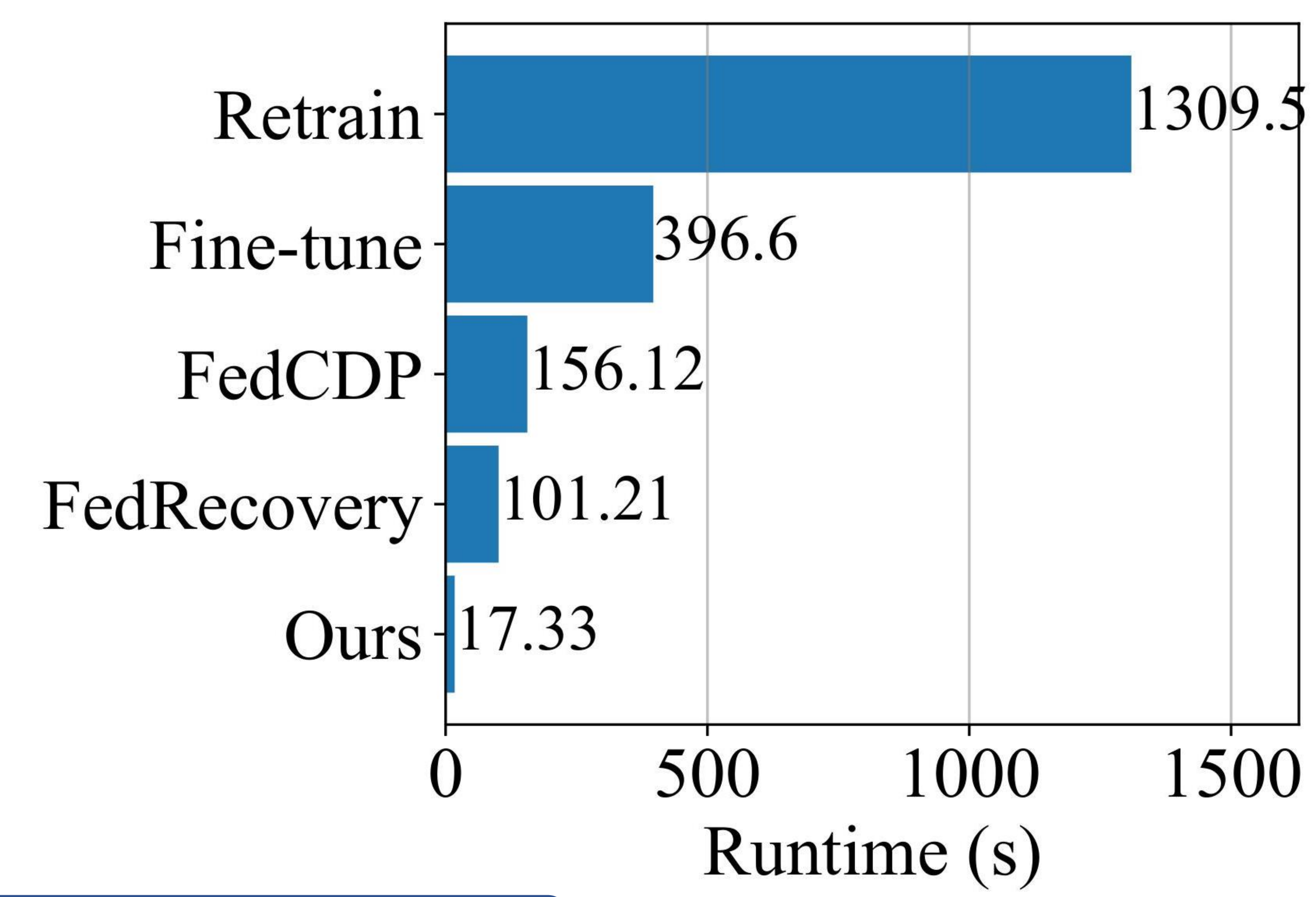
Feature Sensitivity Metrics $s = E_{\delta_F} \frac{\|f(x)+f(x+\delta_F)\|_2}{\|\delta_F\|_2}$

Sensitivity-Guided Optimization $\theta^u = \underset{\theta}{\operatorname{argmin}} E_{(x,y)\in D_u} \frac{1}{N}\sum_{i=1}^{N} \frac{\|f_\theta(x)+f_\theta(x+\delta_{F,i})\|_2}{\|\delta_{F,i}\|_2}$

Utility

Scenario	Dataset	Test Accuracy (%) Baseline	Retrain	FT	Fed CDP	Fed Recovery	Ours
Sensitive	CelebA	94.87	79.46	62.79	34.03	29.78	**92.26**
	Adult	82.45	65.27	61.02	30.19	27.89	**81.02**
Backdoor	CIFAR-100	54.13	73.12	**73.59**	34.30	15.21	69.30
	ImageNet	52.86	67.18	**67.52**	31.17	12.75	65.36
Biased	CMNIST	81.72	98.49	82.54	27.56	25.05	**83.85**
	CelebA	87.35	95.87	88.93	16.98	20.19	**94.62**

Time Efficiency



Effectiveness

Scenario	Metric	Dataset		Baseline	Retrain	FT	FedCDP	FedRecovery	Ours
Sensitive	Model Inversion Attack	CelebA		84.36	47.52	77.43	75.36	71.52	**51.28**
		Adult		87.54	49.28	83.45	72.83	80.39	**49.58**
	Feature Sensitivity	CelebA		0.96	0.07	0.79	0.93	0.91	**0.09**
		Adult		1.31	0.02	0.94	1.07	1.14	**0.05**
Backdoor	Accuracy	CIFAR-100	D_r	54.14	73.54	**73.66**	34.62	15.62	69.57
			D_u	88.98	0.00	65.38	57.29	46.17	**0.15**
		ImageNet	D_r	52.35	67.05	**67.34**	29.74	13.46	65.74
			D_u	83.16	0.00	71.48	62.39	54.92	**0.09**
Biased	Accuracy	CMNIST	D_r	64.94	98.76	67.15	25.85	23.92	**84.31**
			D_u	98.88	98.44	97.95	30.17	27.64	**84.62**
		CelebA	D_r	79.46	96.47	84.45	14.29	16.34	**94.18**
			D_u	96.38	96.11	94.23	21.58	25.72	**94.79**

Conclusion

- Ferrari is a federated feature unlearning framework that efficiently removes sensitive, backdoor, and biased features by requiring only the requesting client's participation. It leverages Lipschitz continuity to reduce model sensitivity and ensure fairness.
- Ferrari preserves privacy, complies with regulatory data deletion requirements, and maintains model performance, making it a practical solution for federated learning environments.