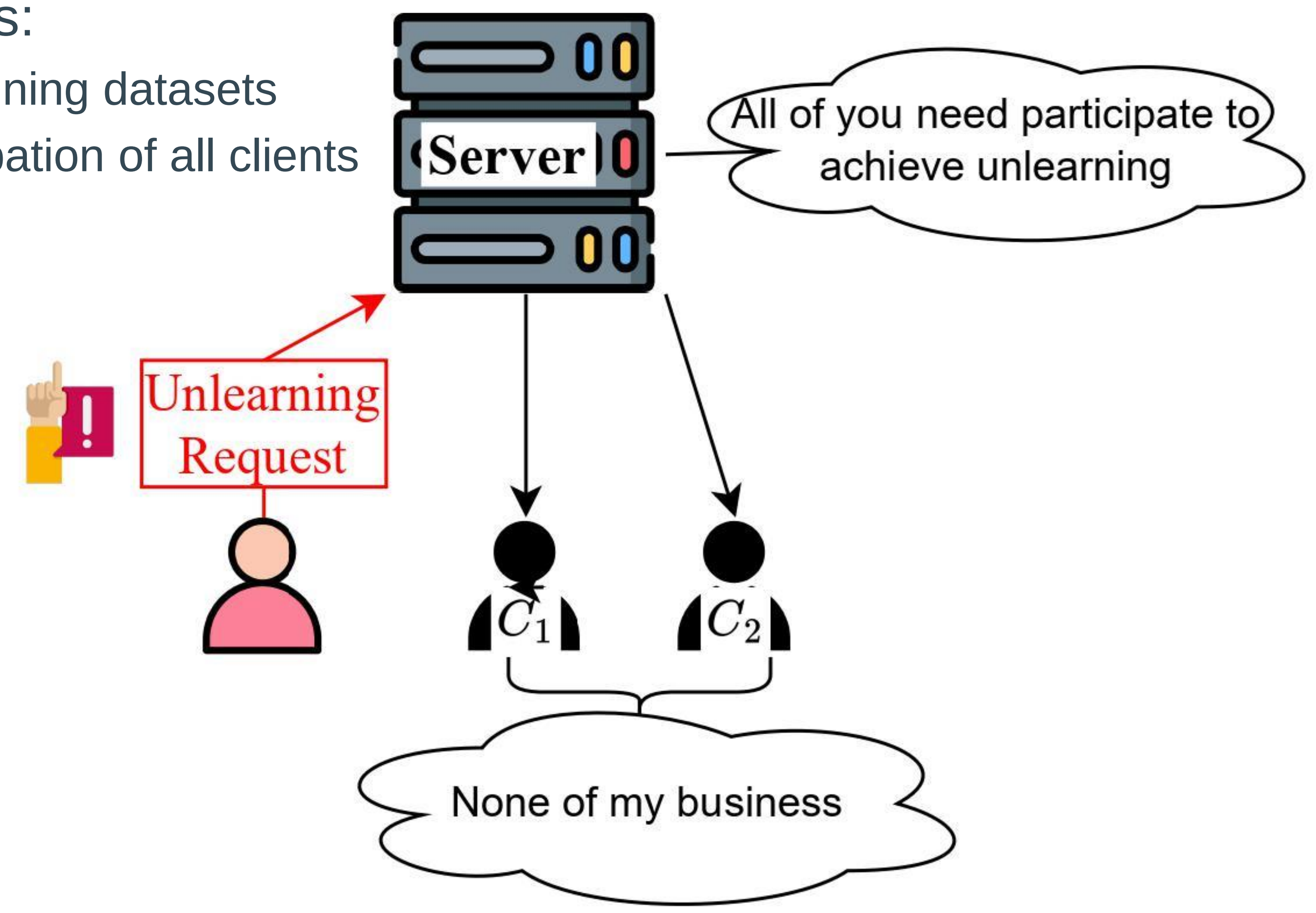
Ferrari: Federated Feature Unlearning via Optimizing Feature Sensitivity

Hanlin Gu[2*], Win Kent Ong[1*], Chee Seng Chan[1], Lixin Fan[2]

Universiti Malaya[1], WeBank AI Lab[2]

Introduction

- Centralized feature unlearning methods impractical in federated settings:
  - Full training datasets
  - Participation of all clients



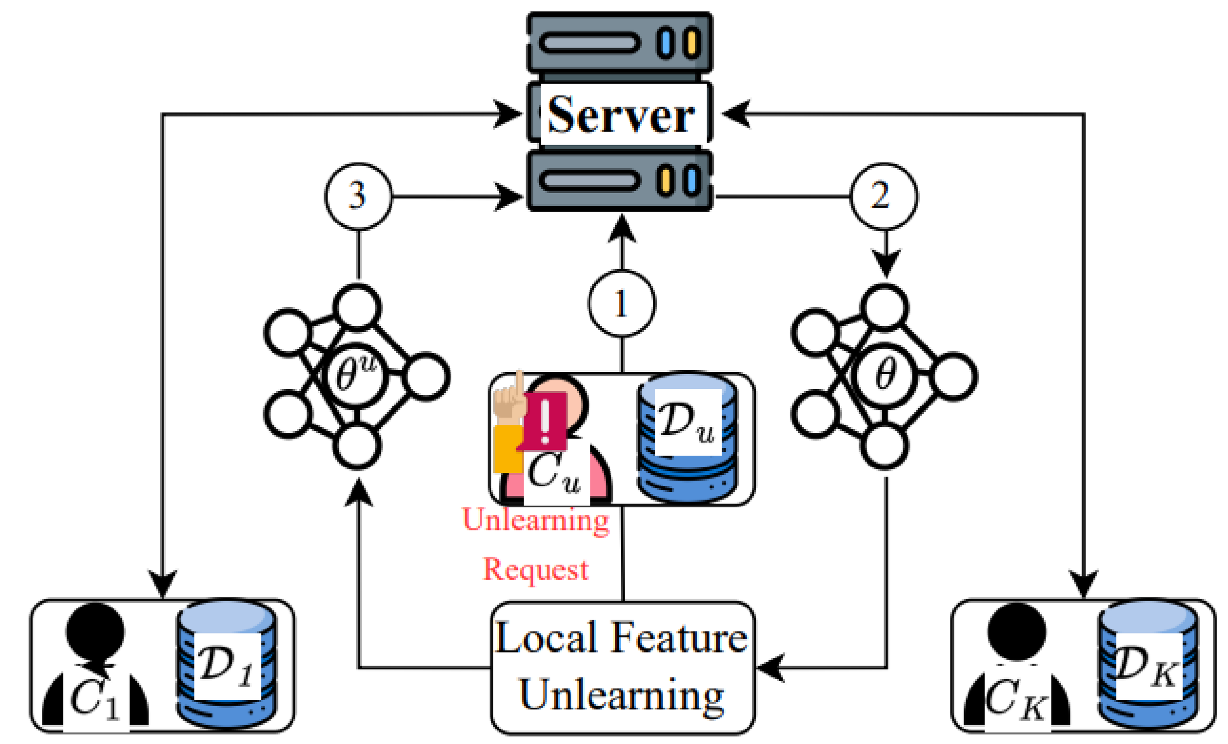
- Difficulty in evaluating the effectiveness of feature unlearning.
  - Conventional method compared to the retrained model without the target feature reduced model utility.

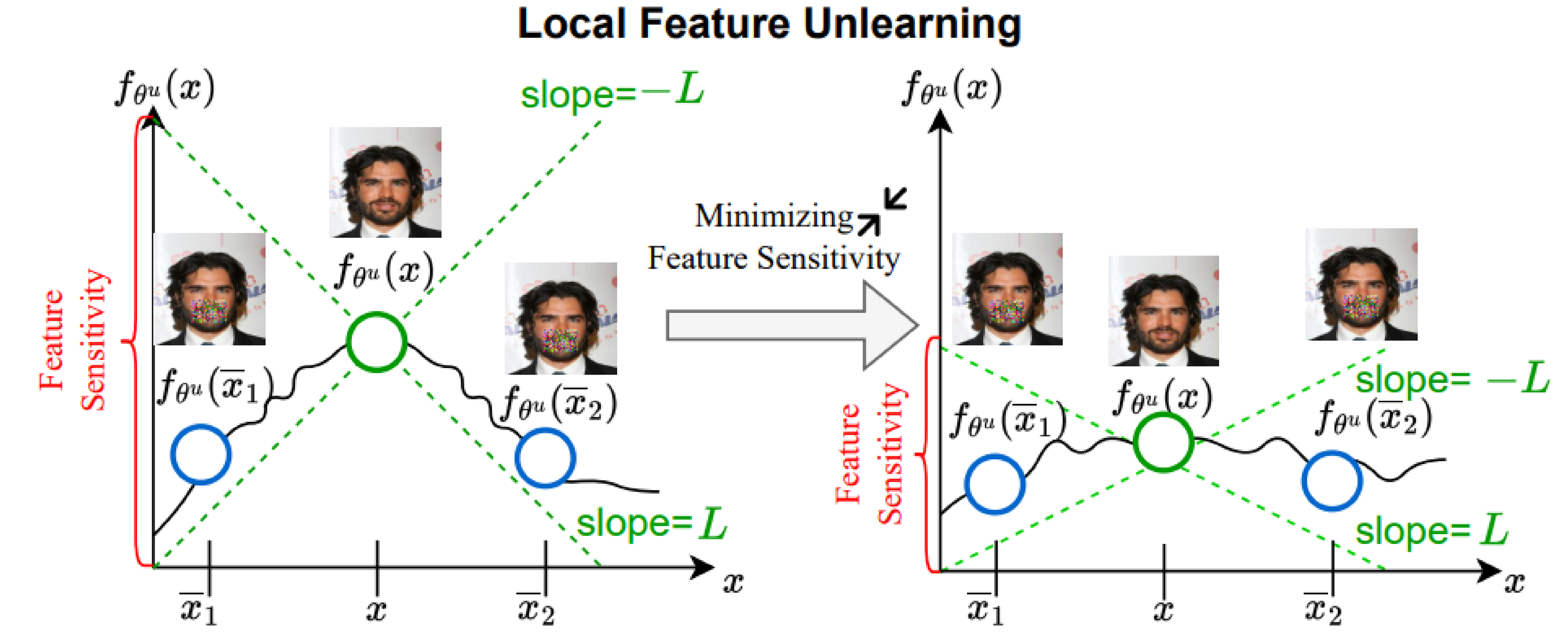
x Acc 95.86% | $\bar{x}_G$ 75.51% | $\bar{x}_B$ 68.37%

Contributions

1. We define the **feature sensitivity** based on Lipschitz continuity and introduce this metric in federated feature unlearning.
2. 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**.
3. We provide theoretical proof and extensive experimental results demonstrate the state-of-the-art **utility** and **effectiveness** of our proposed framework.

Method





- Feature Sensitivity Metrics

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

- Local Unlearning Feature Sensitivity-Guided Optimization

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

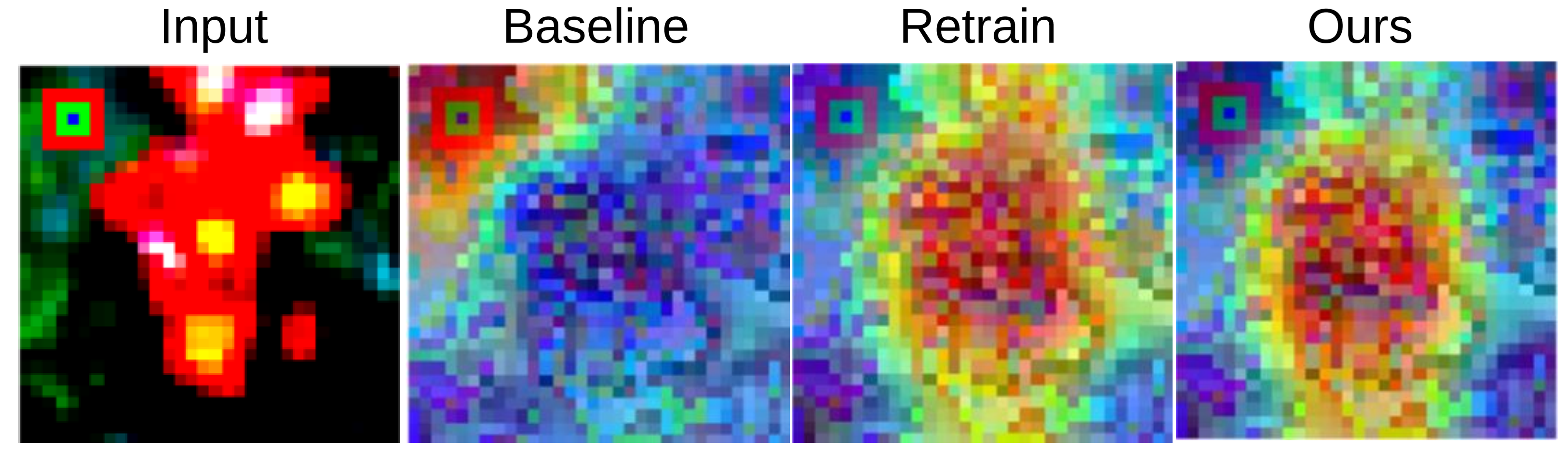
Qualitative Results

- Sensitive Feature Unlearning – Model Inversion Attack

Target | Baseline | Retrain | Ours

- Backdoor Feature Unlearning - GradCAM

Input | Baseline | Retrain | Ours



- Biased Feature Unlearning - GradCAM

Input | Baseline | Retrain | Ours

Quantitative Evaluation

- Sensitive Feature Unlearning

Scenario	Datasets	Unlearn Feature	Feature Sensitivity					
			Baseline	Retrain	Fine-tune	FedCDP [65]	FedRecovery [61]	Ferrari (Ours)
Sensitive	CelebA	Mouth	0.96 $\pm1.41\times10^{-2}$	0.07 $\pm8.06\times10^{-4}$	0.79 $\pm2.05\times10^{-2}$	0.93 $\pm2.87\times10^{-2}$	0.91 $\pm3.41\times10^{-2}$	**0.09 $\pm$3.04**$\times10^{-4}$
	Adult	Marriage	1.31 $\pm1.53\times10^{-2}$	0.02 $\pm6.47\times10^{-4}$	0.94 $\pm6.81\times10^{-2}$	1.07 $\pm7.43\times10^{-2}$	1.14 $\pm2.57\times10^{-2}$	**0.05 $\pm$1.72**$\times10^{-4}$
	Diabetes	Pregnancies	1.52 $\pm0.91\times10^{-2}$	0.05 $\pm5.07\times10^{-4}$	0.96 $\pm1.28\times10^{-2}$	1.23 $\pm3.82\times10^{-2}$	0.83 $\pm5.08\times10^{-2}$	**0.07 $\pm$1.07**$\times10^{-4}$
	IMDB	Names	0.85 $\pm1.07\times10^{-2}$	0.07 $\pm5.38\times10^{-4}$	0.74 $\pm3.81\times10^{-2}$	0.81 $\pm3.27\times10^{-2}$	0.78 $\pm2.41\times10^{-2}$	**0.08 $\pm$1.32**$\times10^{-4}$

Scenario	Datasets	Unlearn Feature	Attack Success Rate(ASR) (%)					
			Baseline	Retrain	Fine-tune	FedCDP [65]	FedRecovery [61]	Ferrari(Ours)
Sensitive	CelebA	Mouth	84.36 ±3.22	47.52 ±1.04	77.43 ±10.98	75.36 ±9.31	71.52 ±6.07	**51.28 ±2.41**
	Adult	Marriage	87.54 ±13.89	49.28 ±2.13	83.45 ±8.44	72.83 ±5.18	80.39 ±10.68	**49.58 ±1.38**
	Diabetes	Pregnancies	92.31 ±7.55	38.89 ±2.52	88.46 ±5.01	81.91 ±8.17	78.27 ±2.47	**42.61 ±1.81**
	IMDB	Names	90.28 ±2.49	40.29 ±1.59	86.74 ±3.81	83.67 ±4.59	80.95 ±3.51	**43.75 ±1.86**

- Backdoor and Biased Feature Unlearning

Scenarios	Datasets	Unlearn Feature		Accuracy (%)					
				Baseline	Retrain	Fine-tune	FedCDP[65]	FedRecovery[61]	Ferrari(Ours)
Backdoor	MNIST	Backdoor pixel-pattern	$\mathcal{D}_r$	95.65 ±1.39	97.19 ±2.49	**96.16 ±0.37**	65.82 ±6.85	40.81 ±4.31	95.93 ±0.45
			$\mathcal{D}_u$	97.43 ±3.69	0.00 ±0.00	72.64 ±0.24	69.37 ±0.83	53.72 ±3.14	**0.11 ±0.01**
	FMNIST		$\mathcal{D}_r$	91.07 ±0.54	93.85 ±1.08	**94.36 ±1.98**	68.46 ±3.39	42.93 ±2.50	92.83 ±0.61
			$\mathcal{D}_u$	94.51 ±6.29	0.00 ±0.00	43.91 ±0.28	72.19 ±0.49	48.15 ±4.37	**0.90 ±0.03**
	CIFAR-10		$\mathcal{D}_r$	87.63 ±1.16	91.12 ±1.60	**92.02 ±3.15**	54.91 ±6.91	27.49 ±4.96	89.91 ±0.95
			$\mathcal{D}_u$	95.05 ±2.30	0.00 ±0.00	88.44 ±0.92	62.75 ±5.07	49.26 ±2.23	**0.29 ±0.04**
	CIFAR-20		$\mathcal{D}_r$	75.06 ±6.41	81.91 ±4.68	**82.67 ±1.32**	55.67 ±6.35	23.76 ±2.17	78.29 ±3.12
			$\mathcal{D}_u$	94.21 ±4.11	0.00 ±0.00	86.53 ±1.47	50.17 ±9.11	50.38 ±4.25	**0.78 ±0.08**
	CIFAR-100		$\mathcal{D}_r$	54.14 ±3.96	73.54 ±5.70	**73.66 ±6.57**	34.62 ±2.24	15.62 ±7.78	69.57 ±3.81
			$\mathcal{D}_u$	88.98 ±6.63	0.00 ±0.00	65.38 ±4.76	57.29 ±3.62	46.17 ±9.25	**0.15 ±0.01**
	ImageNet		$\mathcal{D}_r$	52.35 ±2.25	67.05 ±1.29	**67.34 ±2.73**	29.74 ±4.72	13.46 ±6.53	65.74 ±1.32
			$\mathcal{D}_u$	83.16 ±3.74	0.00 ±0.00	71.48 ±3.69	62.39 ±3.05	54.92 ±5.59	**0.09 ±0.02**
Biased	CMNIST	Color	$\mathcal{D}_r$	64.94 ±7.88	98.76 ±3.65	67.15 ±2.60	25.85 ±1.58	23.92 ±1.08	**84.31 ±2.63**
			$\mathcal{D}_u$	98.88 ±4.90	98.44 ±1.90	97.95 ±1.13	30.17 ±4.69	27.64 ±9.37	**84.62 ±3.59**
	CelebA	Mouth	$\mathcal{D}_r$	79.46 ±2.09	96.47 ±6.15	84.45 ±1.48	14.29 ±0.81	16.34 ±3.43	**94.18 ±3.08**
			$\mathcal{D}_u$	96.38 ±3.87	96.11 ±2.17	94.23 ±0.66	21.58 ±3.48	25.72 ±8.02	**94.79 ±1.48**

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.