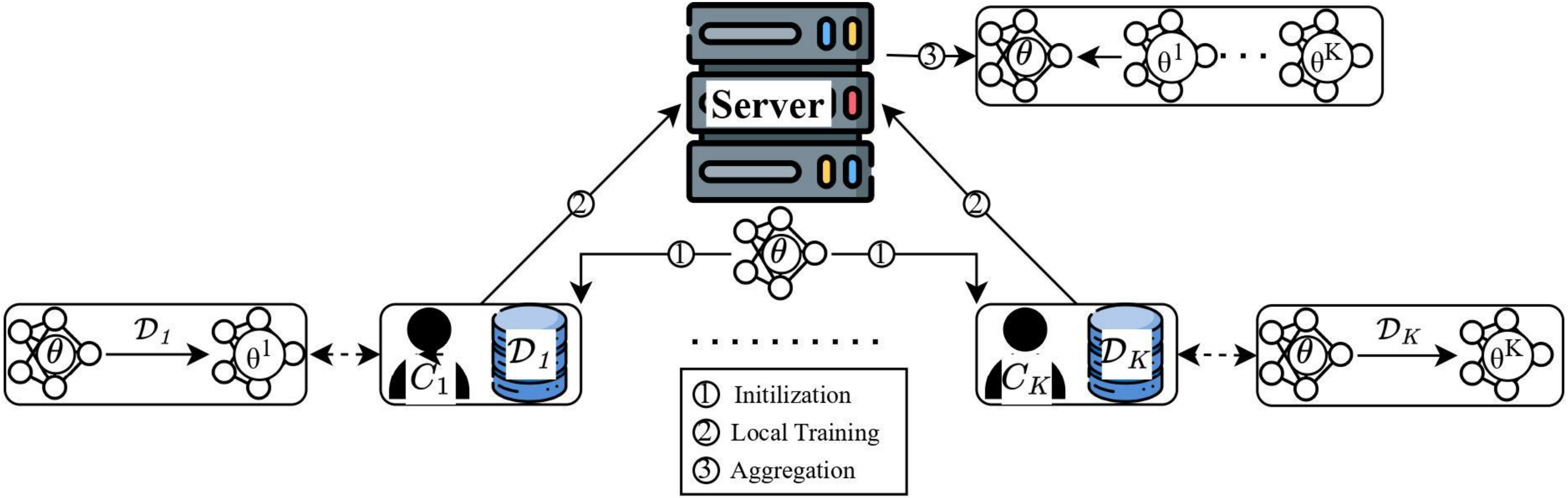
Ferrari: Federated Feature Unlearning via Optimizing Feature Sensitivity

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

[1]Center of Image and Signal Processing, Universiti Malaya
[2]WeBank AI Lab, Shenzhen, China

Introduction – Federated Learning

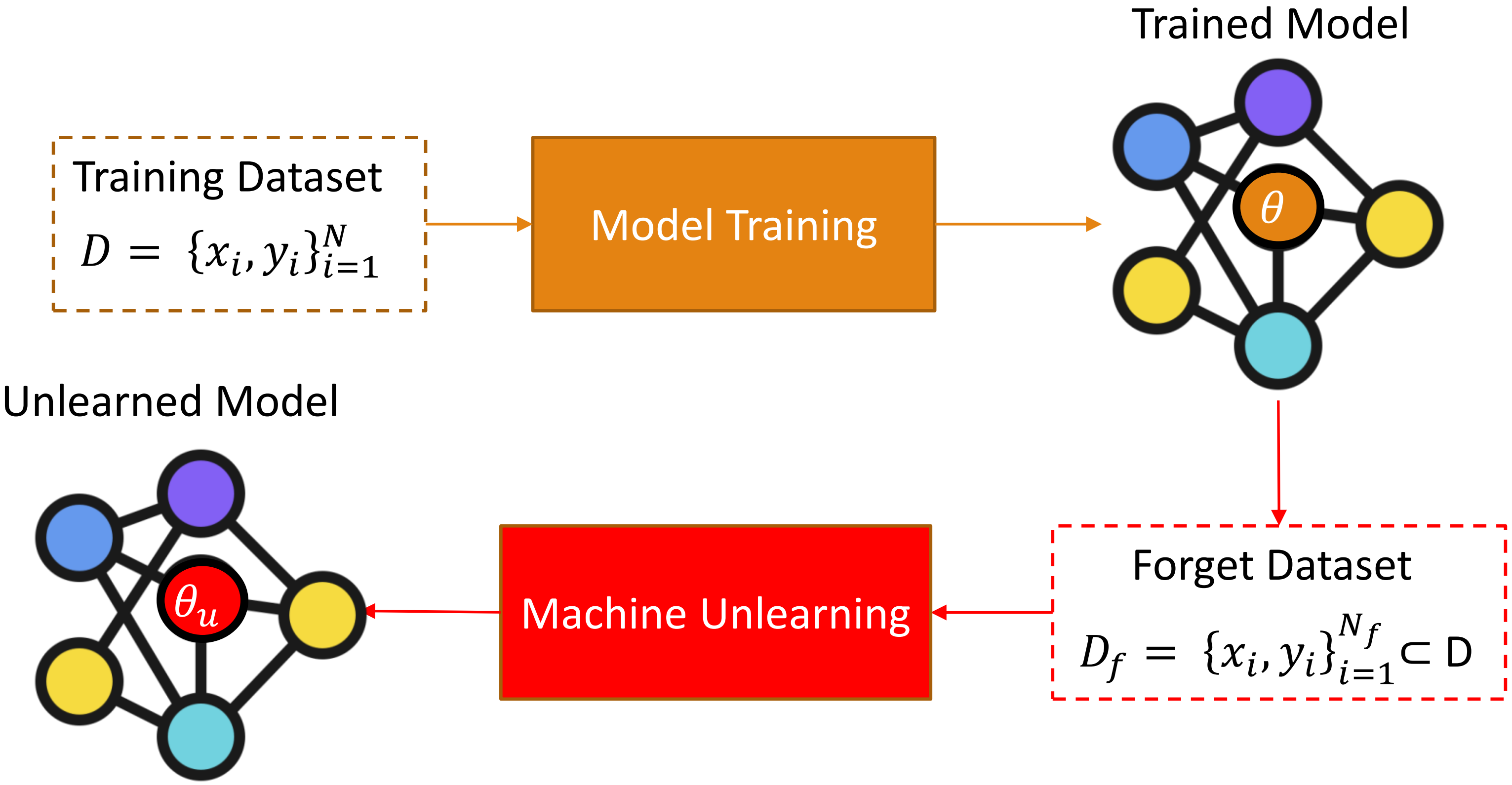


Machine Learning algorithm enables multiple parties to collaboratively train a model

- Without sharing private data, only sharing trained weights
- Better data privacy protection, reducing the risk of privacy leakage

Introduction – Machine Unlearning

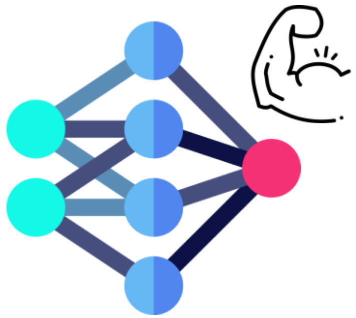
- Remove the influence of a subset of its training dataset from the trained neural network.



Introduction – Machine Unlearning

- PRIVACY REGULATION LAWS
  - California Consumer Privacy Act (CCPA)
  - General Data Protection Regulation (GDPR)
  - Consumer Privacy Protection Act (CPPA)
  - Secure the right to be forgotten

- REMOVE OUTDATED OR MISLABELLED TRAINING DATA
  - Improve model robustness



Motivation

1. Federated Unlearning
   - Current works focus on isolated data points
   - Client, sample or class level unlearning
2. Centralized Feature Unlearning
   - Impractical for Federated Learning due to participation of all client (all datasets).
3. Difficulty in evaluating the effectiveness of feature unlearning.
   - Conventional method compared to the retrained model without the target feature **reduced model utility**.

Contributions

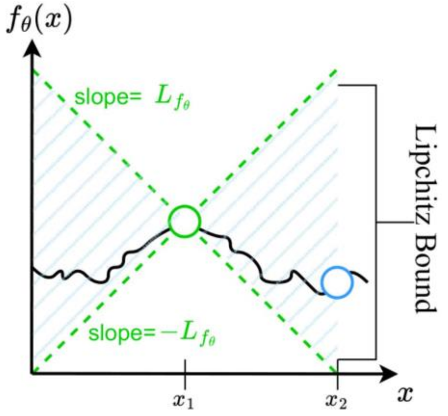
I. We define the Feature Sensitivity metric based on Lipschitz Continuity

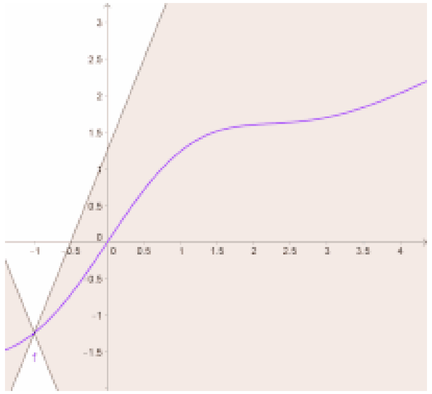
II. We proposed an effective **federated feature unlearning** framework
- allowing clients to selectively unlearn specific features
- without the participation of other clients
- optimizing feature sensitivity locally

III. We provide theoretical proof and extensive experimental results demonstrate the state-of-the-art **utility** and **effectiveness** of our proposed framework.

Revisit - Lipschitz Continuity

Lipschitz continuity quantifies the sensitivity of a function, by quantifying how function values change with respect to variations in the independent variable





Exist a non-negative Lipschitz constant

$$\underbrace{||f_\theta(x_1) - f_\theta(x_2)||_Y}_{\text{Output}} = L_{f_\theta}\underbrace{||x_1 - x_2||_X}_{\text{Input}}, \forall(x_1, x_2) \in \mathcal{X}$$

$$\sup_{x_1, x_2 \in \mathcal{X}, x_1 \neq x_2} \frac{||f_\theta(x_1) - f_\theta(x_2)||_Y}{||x_1 - x_2||_X} \leq L_{f_\theta}$$

Bounded Rate of Change - Average rate of change of the function bounded by Lipschitz bound.

$$-L_{f_\theta} \leq \frac{||f_\theta(x_1) - f_\theta(x_2)||_Y}{||x_1 - x_2||_X} \leq L_{f_\theta}$$

Feature Sensitivity

Feature Sensitivity: $s = \frac{\|f(x)-f(\bar{x})\|}{\|(x)-(\bar{x})\|}$

$$s = \frac{\|f(x)-f(x+\delta)\|}{\|(x)-(x+\delta)\|}$$

$$s = \frac{\|f(x)-f(x+\delta)\|}{\|\delta\|}$$

$x =$

$\bar{x} = x + \delta =$

Intuition Sensitivity-Guided Optimization

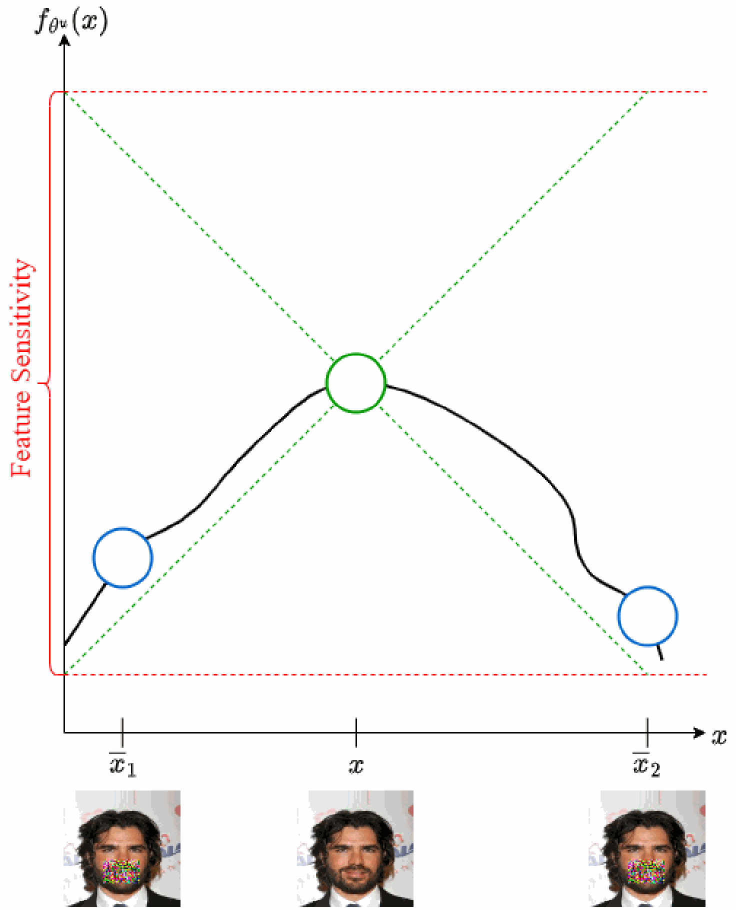
Core Idea: Optimize Feature Sensitivity via Guided Lipschitz Bound

$$\mathcal{L} = \frac{\|f(x)-f(x+\delta)\|}{\|\delta\|}, (x, y) \in D_u$$

Feature Sensitivity as guided loss function to optimize the unlearn model θ^u via gradient descent

$$\theta^u \leftarrow \theta^u - \eta \cdot \nabla_{\theta^u}(\mathcal{L})$$

$$\nabla_{\theta^u}(\mathcal{L}) = \frac{\partial \mathcal{L}}{\partial \theta_u}$$



Theoretical Proof – Utility Loss

$$\ell_1 = \min_{\|\delta_{\mathcal{F}}\| \geq C} \mathbb{E}_{(x,y)\in\mathcal{D}} \min_{\theta} \ell\big(f_\theta(x+\delta_{\mathcal{F}}), y\big)$$

$$\ell_2 = \max_{\|\delta_{\mathcal{F}}\| \leq C} \mathbb{E}_{(x,y)\in\mathcal{D}} \min_{\theta} \ell\big(f_\theta(x+\delta_{\mathcal{F}}), y\big)$$

Assumption 1. *Assume* $\ell_2 \leq \ell_1$

larger perturbations would naturally lead to greater utility loss

Assumption 2. *Suppose the federated model achieves zero training loss.*

Theorem 1. *If Assumption 1 and Assumption 2 hold, the utility loss of unlearned model obtained by Algorithm 1 is less than the utility loss with unlearning successfully, i.e.*

$$\boxed{\ell_u \leq \ell_1,} \tag{3.10}$$

where $\ell_u = \mathbb{E}_{(x,y)\in\mathcal{D}}\ell(f_{\theta^u}(x), y)$

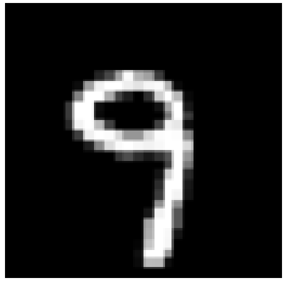
Experimental Setup -Models and Datasets

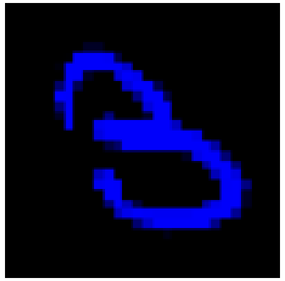
TABULAR DATASET

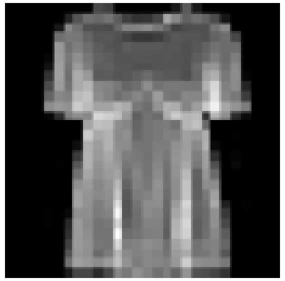
- Fully-Connected Linear Neural Network
- Adult Census Income (Adult) Dataset - includes 48, 842 records with 14 attributes to predict if a person earns over $50K a year based on the census attributes and marital status as the sensitive feature that aim to unlearn.
- Diabetes Dataset: includes 768 personal health to predict if a person has diabetes and number of pregnancies as the sensitive feature that aim to unlearn.

IMAGE DATASET

- ResNet-18 (Convolutional Neural Networks)

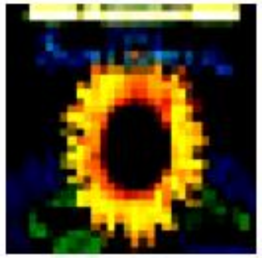

MNIST


CMNIST


FMNIST


CIFAR-10


CIFAR-20


CIFAR-100


CelebA

Experimental Setup - Baselines

- Baseline – Original model before unlearning
- Retrain – Model training without the presence of unlearn feature
- Fine-tune – Fine-tuning baseline model with the retain dataset.
- FedCDP - A Federated Unlearning framework that achieves class unlearning by utilizing Term Frequency Inverse Document Frequency (TF-IDF) guided channel pruning, which selectively removes the most discriminative channels related to the target category and followed by fine-tuning without retraining from scratch.
- FedRecovery - A Federated Unlearning framework that achieves client unlearning by removing the influence of a client’s data from the global model using a differentially private machine unlearning algorithm that leverages historical gradient submissions without the need for retraining

Effectiveness - Sensitive Feature Unlearning

Model Inversion Attack – Attack Success Rate

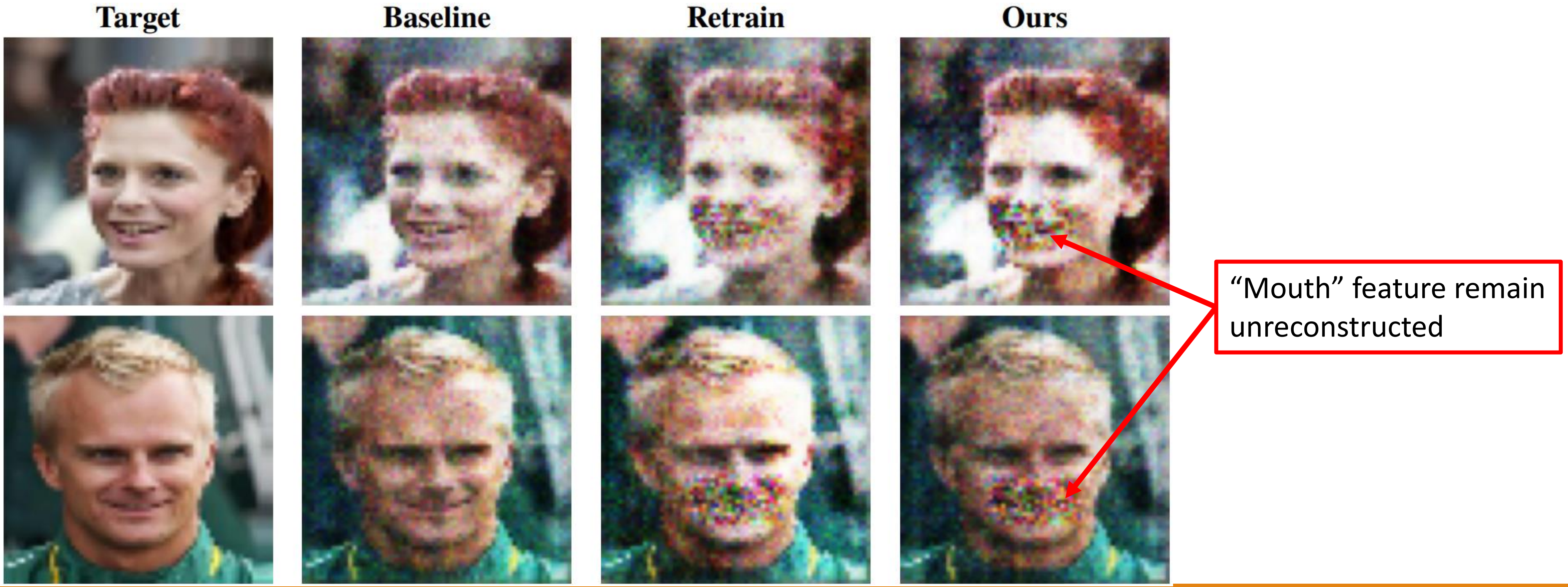
Scenario	Datasets	Unlearn Feature	Attack Success Rate(ASR) (%) ↓					
			Baseline	Retrain	Fine-tune	FedCDP	FedRecovery	Ours
Sensitive	**CelebA**	Mouth	84.36 ±3.22	47.52 ±1.04	77.43 ±10.98	75.36 ±9.31	71.52 ±6.07	**51.28 ±2.41**
	Adult	Marriage	87.54 ±13.89	49.28 ±2.13	83.45 ±8.44	72.83 ±5.18	80.39 ±10.68	**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**

Feature Sensitivity

Scenario	Datasets	Unlearn Feature	Feature Sensitivity					
			Baseline	Retrain	Fine-tune	FedCDP	FedRecovery	Ours
Sensitive	**CelebA**	Mouth	0.96 $\pm 1.41\times10^{-2}$	0.07 $\pm 8.06\times10^{-4}$	0.79 $\pm 2.05\times10^{-2}$	0.93 $\pm 2.87\times10^{-2}$	0.91$\pm 3.41\times10^{-2}$	**0.09 $\pm 3.04\times10^{-4}$**
	Adult	Marriage	1.31 $\pm 1.53\times10^{-2}$	0.02 $\pm 6.47\times10^{-4}$	0.94 $\pm 6.81\times10^{-2}$	1.07 $\pm 7.43\times10^{-2}$	1.14 $\pm 2.57\times10^{-2}$	**0.05 $\pm 1.72\times10^{-4}$**
	Diabetes	Pregnancies	1.52 $\pm 0.91\times10^{-2}$	0.05 $\pm 5.07\times10^{-4}$	0.96 $\pm 1.28\times10^{-2}$	1.23 $\pm 3.82\times10^{-2}$	0.83 $\pm 5.08\times10^{-2}$	**0.07 $\pm 1.07\times10^{-4}$**

Effectiveness - Sensitive Feature Unlearning

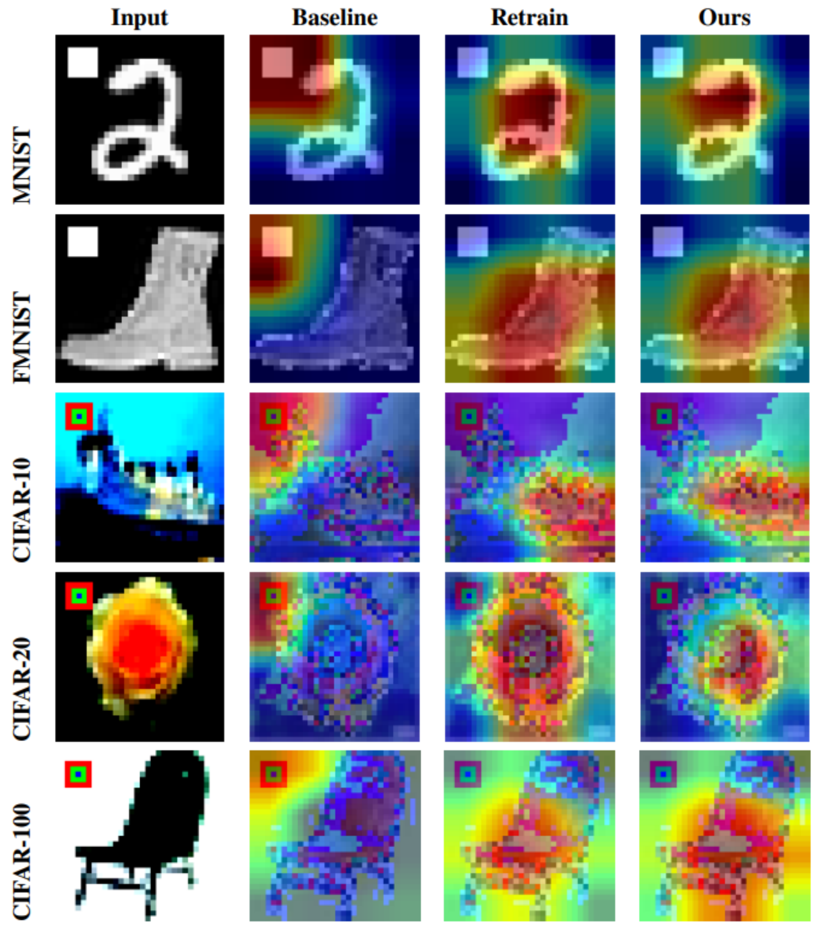
Model Inversion Attack – Reconstructed Images

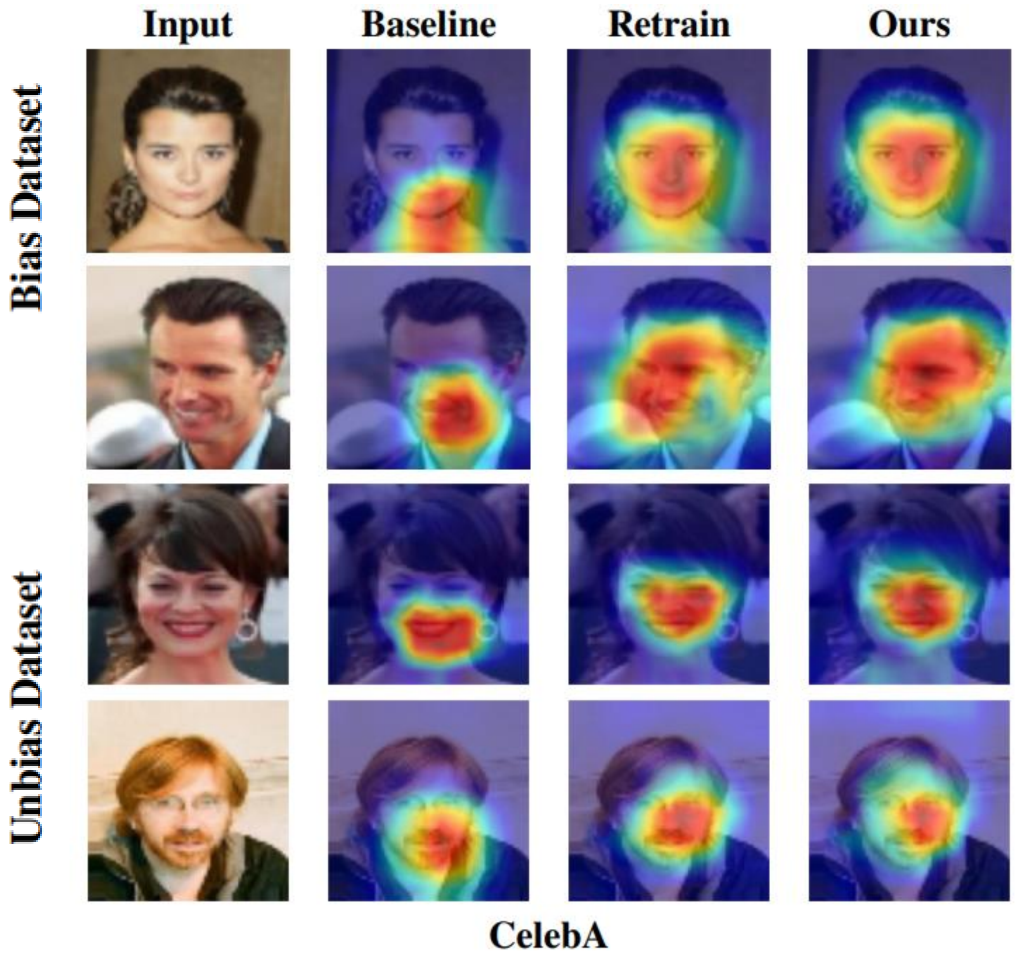


Effectiveness - Backdoor & Biased Feature Unlearning

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**

Effectiveness - Backdoor & Biased Feature Unlearning





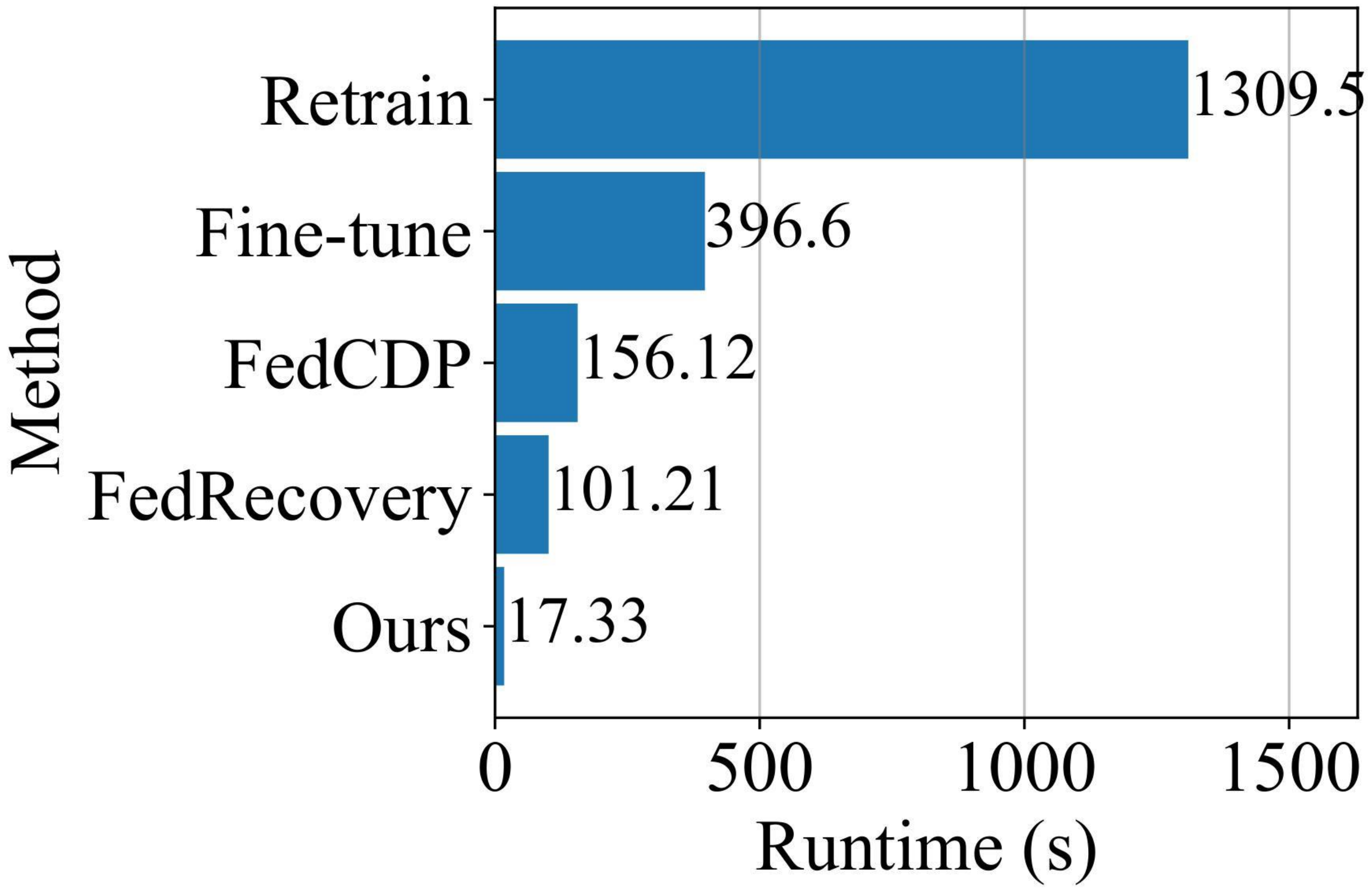
Backdoor Feature

Biased Feature

Utility

Scenarios	Datasets	Unlearn Feature	Accuracy(%)↑					
			Baseline	**Retrain**	**Fine-tune**	**FedCDP[65]**	**FedRecovery[61]**	**Ferrari (Ours)**
Sensitive	**CelebA**	Mouth	94.87 ±1.38	79.46 ±2.32	62.79 ±1.62	34.03 ±4.20	29.78 ±6.69	**92.26 ±1.73**
	Adult	Marriage	82.45 ±2.59	65.27 ±0.58	61.02 ±1.05	30.19 ±1.62	27.89 ±3.71	**81.02 ±0.58**
	Diabetes	Pregnancies	82.11 ±0.49	64.19 ±0.72	59.57 ±0.68	36.71 ±4.56	17.56 ±2.32	**79.53 ±0.79**
	IMDB	Names	91.39 ±1.57	83.27 ±2.05	72.15 ±1.92	48.36 ±2.79	37.93 ±2.84	**89.15 ±1.32**
Backdoor	**MNIST**	Backdoor Pixel Pattern	94.75 ±4.88	96.23 ±0.16	**96.85 ±0.91**	65.31 ±4.39	40.52 ±7.38	95.83 ±1.14
	FMNIST		90.68 ±2.19	92.98 ±0.75	**93.52 ±1.63**	67.62 ±0.81	42.24 ±4.45	92.61 ±1.57
	CIFAR-10		87.55 ±3.71	90.92 ±1.83	**91.23 ±0.44**	53.98 ±2.17	27.16 ±9.68	89.52 ±2.18
	CIFAR-20		74.47 ±2.38	81.61 ±1.75	**82.52 ±0.69**	54.76 ±0.98	23.02 ±3.11	78.34 ±2.35
	CIFAR-100		54.13 ±7.62	73.12 ±1.54	**73.59 ±1.66**	34.30 ±0.42	15.21 ±5.83	69.30 ±2.27
	ImageNet		52.86 ±4.14	67.18 ±2.07	**67.52 ±1.69**	31.17 ±3.96	12.75 ±5.27	65.36 ±1.84
Biased	**CMNIST**	Color	81.72 ±3.41	98.49 ±1.46	82.54 ±0.78	27.56 ±1.71	25.05 ±5.09	**83.85 ±1.63**
	CelebA	Mouth	87.35 ±4.07	95.87 ±1.52	88.93 ±2.65	16.98 ±0.23	20.19 ±7.21	**94.62 ±2.49**

Time Efficiency



Conclusion

- To best of our knowledge, this is the first work to achieve feature unlearning within Federated Learning settings.
- The proposed Federated Feature Unlearning framework effectively achieves feature unlearning via the proposed Sensitivity-Guided Optimization algorithm.
- Theoretical analysis and experimental results, both quantitative and qualitatively.
- Practical Federated Feature Unlearning Framework without participation of all clients, only participation of unlearn client is needed.

Thank you for listening!

Paper

Code

Email