Weekly Meeting

An Information Theoretic Approach to Machine Unlearning

Foster et al.,
University of Cambridge

Published in Transactions on Machine Learning Research (03/2025)

2025.07.22

SecAI Lab

SUNGKYUNKWAN UNIVERSITY

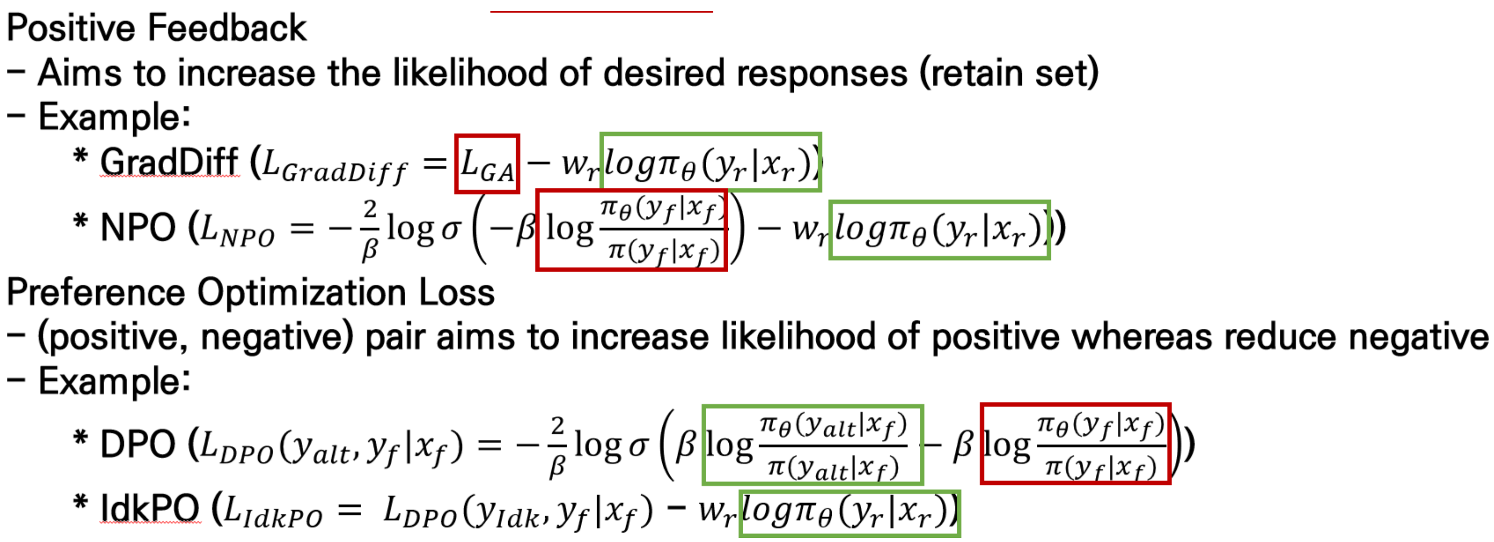
Machine Unlearning?

- GDPR (Right to be forgotten)
  - DB: Just delete user's info
  - Deep-learned model: challenge

- Machine Unlearning: open problem with two desiderata
  - Remove the influence of the selected subset of data (i.e., forget set)
  - Maintain model performance on retained data (i.e., retain set)

Traditional Unlearning vs. Zero-shot Unlearning

- Traditional unlearning
  - Strong assumptions: access to all (or, subset of) training dataset (red: forget; green: retain set)



Positive Feedback

– Aims to increase the likelihood of desired responses (retain set)

– Example:

* GradDiff ($L_{GradDiff} = \boxed{L_{GA}} - w_r \boxed{log\pi_\theta(y_r|x_r)}$)

* NPO ($L_{NPO} = -\frac{2}{\beta}\log\sigma\left(-\beta \boxed{\log\frac{\pi_\theta(y_f|x_f)}{\pi(y_f|x_f)}}\right) - w_r \boxed{log\pi_\theta(y_r|x_r)}$)

Preference Optimization Loss

– (positive, negative) pair aims to increase likelihood of positive whereas reduce negative

– Example:

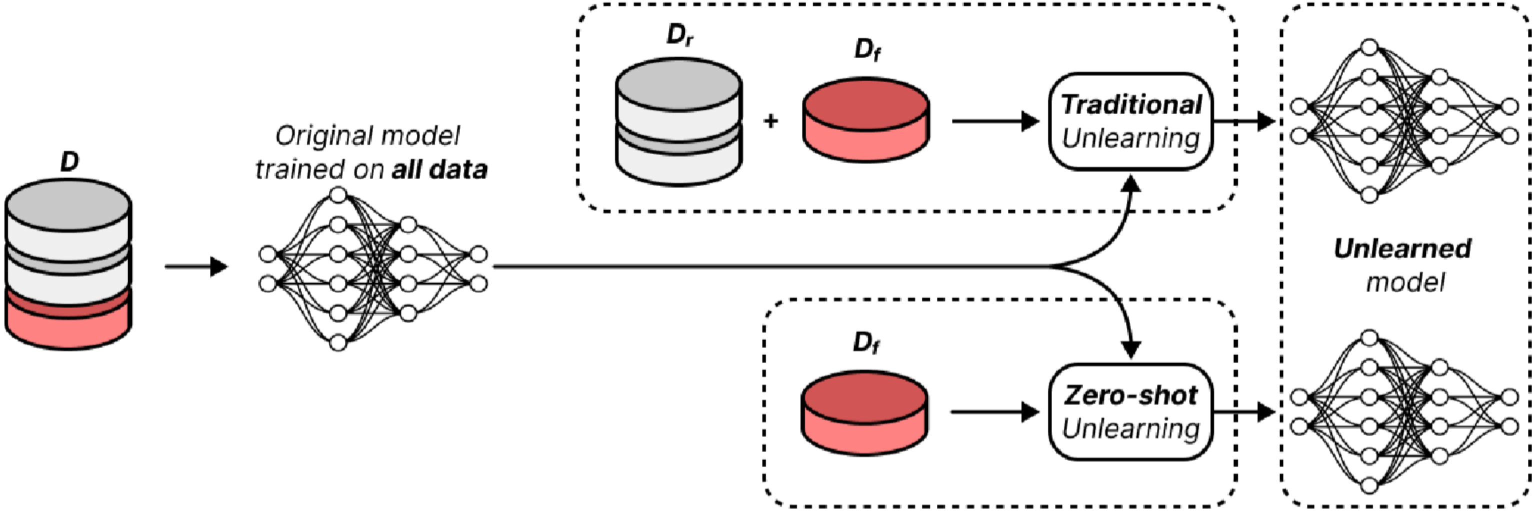
* DPO ($L_{DPO}(y_{alt}, y_f|x_f) = -\frac{2}{\beta}\log\sigma\left(\beta \boxed{\log\frac{\pi_\theta(y_{alt}|x_f)}{\pi(y_{alt}|x_f)}} - \beta \boxed{\log\frac{\pi_\theta(y_f|x_f)}{\pi(y_f|x_f)}}\right)$)

* IdkPO ($L_{IdkPO} = L_{DPO}(y_{Idk}, y_f|x_f) - w_r \boxed{log\pi_\theta(y_r|x_r)}$)

➔ In reality, training set is often available (cost of storage, limited duration access to datasets)

Traditional Unlearning vs. Zero-shot Unlearning

- Zero-shot unlearning
  - Chundawat [1] proposed, no need for retain set; but limited to unlearning singular “full class”
  - This paper presents zero-shot unlearning of “arbitrary data point”



[1] Vikram S Chundawat, Ayush K Tarun, Murari Mandal, and Mohan Kankanhalli. Zero-shot machine unlearning. IEEE Transactions on Information Forensics and Security, 2023b.

Key Idea

- Information-Theoretic Perspective [1]
  - Low information gain samples: Can be inferred from other data → lie in low-gradient regions → minimal boundary impact when removed
  - High information gain samples: Unique/memorized → lie in high-gradient regions → pronounced effect when removed

➔ Samples with low information gain use generalized knowledge (privacy-safe), while high information gain samples are likely memorized (privacy-infringing).

* This paper focuses on unlearning classification models, not generative models.

[1] Golatkar et al., Eternal sunshine of the spotless net: Selective forgetting in deep networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 9304–9312, 2020.

Brief overview video

Background: Curvature

- Curvature: measuring how rapidly function bends via second order derivative
  - Reflects how sensitive the decision boundary is around data point

$$\kappa = \frac{\|\nabla^2 f(x)\|_F}{\|\nabla f(x)\|}$$

- Typically, well-learned models exhibit sharp decision boundaries with a large rate of change, whereas flatter behavior within the class [1]
  -> High curvature region tend to be decision boundary of model

[1] Fridovich-Keil et al., Spectral bias in practice: The role of function frequency in generalization. Advances in Neural Information Processing Systems, 35:7368–7382, 2022.

Hypothesis, Information Gain Calculation

- Hypothesis
  - Number of forget set: 1 (larger would be easier)
  - Neural network is well-trained and generalizes to in distribution test set well

- Define sample neighborhood: data points (z) within reach of forget set (x) in geometric space

For sample $x \in \mathcal{X}$ with radius $r > 0$:

$$B_r(x) = \{z \in \mathcal{X} : \|z - x\| \leq r\}$$

- Information content of data point x: ratio of data points in sample neighborhood (Br(x)) aligns with forget set's class

$$\alpha(x) = \frac{1}{|B_r(x)|} \sum_{z \in B_r(x)} \mathbb{I}\{c(z) = c(x)\}$$

- Sample classification with threshold τ ∈ [0,1]

Low information: $\alpha(x) \geq \tau$ (can be inferred from neighbors)
High information: $\alpha(x) < \tau$ (unique influence on model)

Proposed JiT (Just in Time) algorithm

- Core loss function: simply, moving toward direction which has same class with aimed forget set (x).
  - Loss would be minimized when f(x) aligns with f(x+noise), whereas moving backward from different class

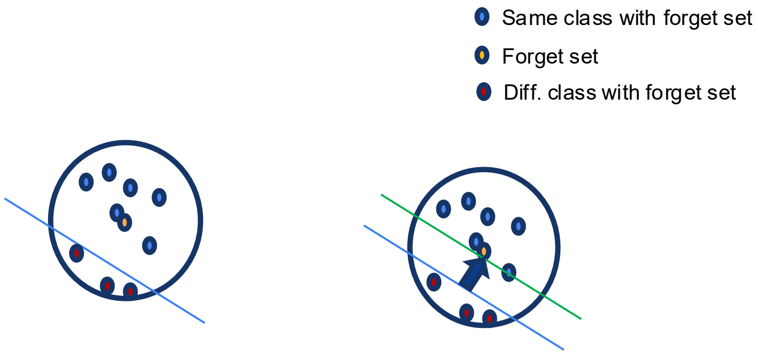
For each forget sample $x \in \mathcal{D}_f$:

$$\ell = \mathbb{E}\left[\frac{\|f_\theta(x) - f_\theta(x+\xi)\|_2^2}{\|\xi\|_2^2}\right]$$

First order approximation with N perturbations:

$$\ell \approx \frac{1}{N}\sum_{j=1}^{N}\frac{\|f_\theta(x) - f_\theta(x+\xi_j)\|_2^2}{\|\xi_j\|_2^2}$$

where $\xi_j \sim \mathcal{N}(0, \sigma^2)$



Before JiT

After JiT
Decision boundary
moves toward blue dots

Geometry of moving boundaries

* Curvature (measuring how rapidly function bends via second order derivative) reflects how sensitive the decision boundary is around data point

- In high curvature region, JiT moves decision boundary toward forget set’s class (1, blue)
  - In low curvature region, boundary shows minimum movement

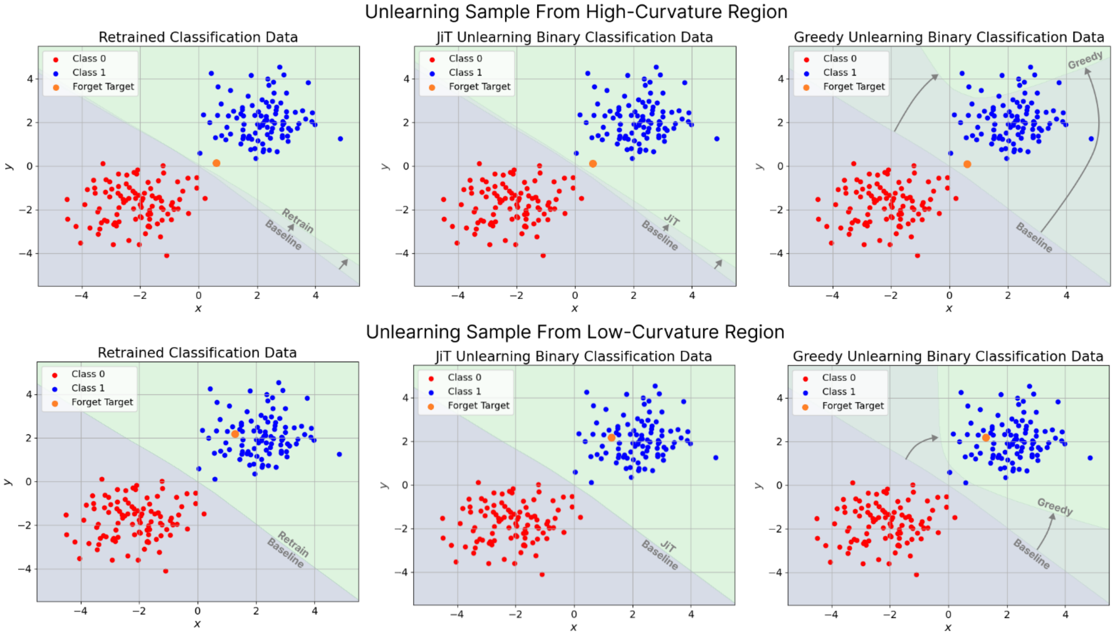
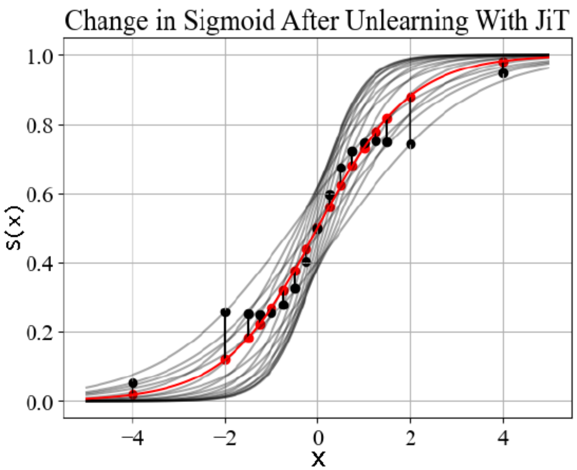


Figure 2: Demonstration of how the boundary of a classifier moves during unlearning. Retrained model is the gold standard. Removing a sample from a low-gradient region has almost no effect on the retrained model, whereas removing a sample from high gradient space has significant impact. In this low-dimensional setting, JiT successfully reconstructs the retrained boundary, whereas naively training to mislabel the forget sample completely destroys the trained model.

Effect of JiT in Unlearned Models – Sigmoid example



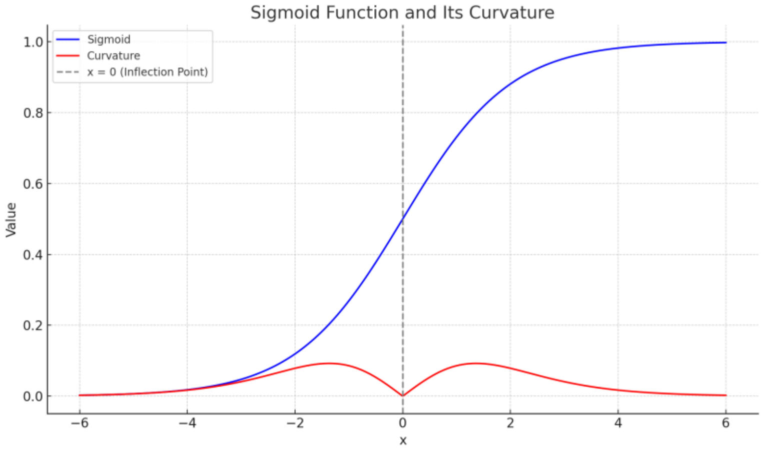


Figure 3: Change in sigmoid after unlearning with JiT. Red dots are unlearnt samples, black dots are the location on the new sigmoid post-JiT.

- Forgetting has minimal impact when the forget sample is in a low-information (flat/low-curvature) region.
- Forgetting samples in high-gradient areas leads to a noticeable deformation of the sigmoid, shifting decision boundaries.

Experimental Setup

- Datasets: Align with [1], used CIFAR-{10, 20, 100}, PinsFaceRecognition(17k+ images, 105 celebrities)
- Model: ResNet18, VisionTransformer, lr=0.1, Adam
- Unlearning tasks:
  - Single (Full)-class forgetting
  - Sub-class forgetting (e.g., ‘rockets’ in vehicle)
  - Random observations forgetting (i.e., uniform sample of training set)
- Comparison models
  - BSLN(baseline): Not unlearned
  - RTRN(Retrained): Retrained only on retain data
  - FNTN(Finetuned for 5 epochs)
  - SSD, GKT, EMMN, SCRUB, BT, AMN, UNSIR: SoTA methods
- Evaluation metrics
  - Retain set accuracy
  - Forget set accuracy
  - MIA Score
  - Runtime

[1] Forster et al., Fast Machine Unlearning Without Retraining Through Selective Synaptic Dampening, AAAI, 2024

Evaluation

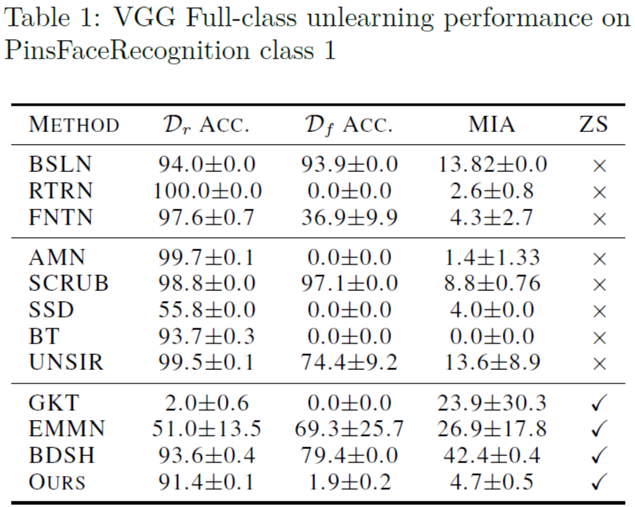


Table 1: VGG Full-class unlearning performance on PinsFaceRecognition class 1

METHOD	$\mathcal{D}_r$ ACC.	$\mathcal{D}_f$ ACC.	MIA	ZS
BSLN	94.0±0.0	93.9±0.0	13.82±0.0	×
RTRN	100.0±0.0	0.0±0.0	2.6±0.8	×
FNTN	97.6±0.7	36.9±9.9	4.3±2.7	×
AMN	99.7±0.1	0.0±0.0	1.4±1.33	×
SCRUB	98.8±0.0	97.1±0.0	8.8±0.76	×
SSD	55.8±0.0	0.0±0.0	4.0±0.0	×
BT	93.7±0.3	0.0±0.0	0.0±0.0	×
UNSIR	99.5±0.1	74.4±9.2	13.6±8.9	×
GKT	2.0±0.6	0.0±0.0	23.9±30.3	✓
EMMN	51.0±13.5	69.3±25.7	26.9±17.8	✓
BDSH	93.6±0.4	79.4±0.0	42.4±0.4	✓
OURS	91.4±0.1	1.9±0.2	4.7±0.5	✓

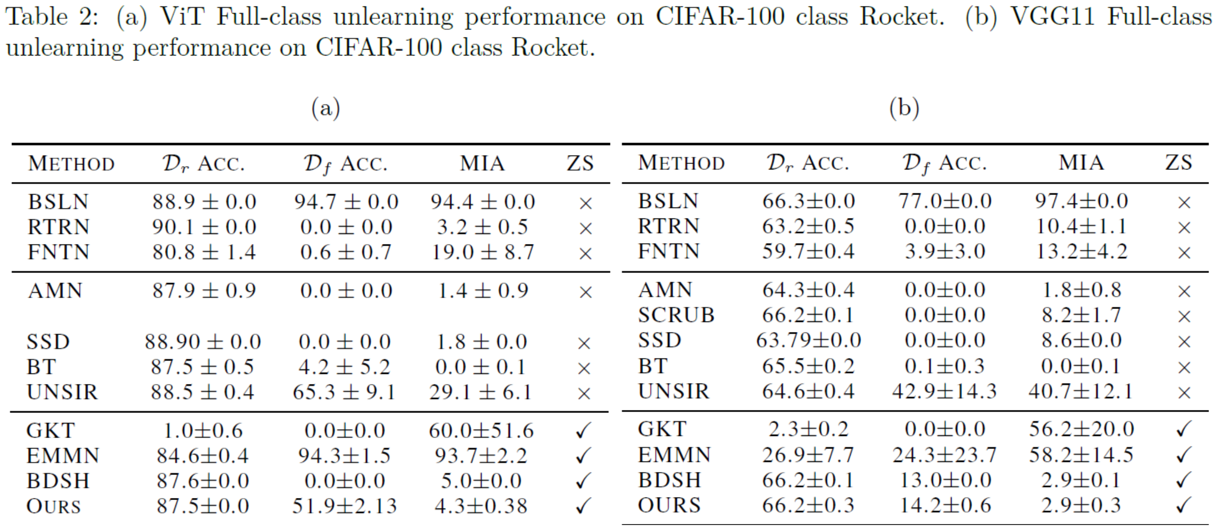


Table 2: (a) ViT Full-class unlearning performance on CIFAR-100 class Rocket. (b) VGG11 Full-class unlearning performance on CIFAR-100 class Rocket.

(a)

METHOD	$\mathcal{D}_r$ ACC.	$\mathcal{D}_f$ ACC.	MIA	ZS
BSLN	88.9 ± 0.0	94.7 ± 0.0	94.4 ± 0.0	×
RTRN	90.1 ± 0.0	0.0 ± 0.0	3.2 ± 0.5	×
FNTN	80.8 ± 1.4	0.6 ± 0.7	19.0 ± 8.7	×
AMN	87.9 ± 0.9	0.0 ± 0.0	1.4 ± 0.9	×
SSD	88.90 ± 0.0	0.0 ± 0.0	1.8 ± 0.0	×
BT	87.5 ± 0.5	4.2 ± 5.2	0.0 ± 0.1	×
UNSIR	88.5 ± 0.4	65.3 ± 9.1	29.1 ± 6.1	×
GKT	1.0±0.6	0.0±0.0	60.0±51.6	✓
EMMN	84.6±0.4	94.3±1.5	93.7±2.2	✓
BDSH	87.6±0.0	0.0±0.0	5.0±0.0	✓
OURS	87.5±0.0	51.9±2.13	4.3±0.38	✓

(b)

METHOD	$\mathcal{D}_r$ ACC.	$\mathcal{D}_f$ ACC.	MIA	ZS
BSLN	66.3±0.0	77.0±0.0	97.4±0.0	×
RTRN	63.2±0.5	0.0±0.0	10.4±1.1	×
FNTN	59.7±0.4	3.9±3.0	13.2±4.2	×
AMN	64.3±0.4	0.0±0.0	1.8±0.8	×
SCRUB	66.2±0.1	0.0±0.0	8.2±1.7	×
SSD	63.79±0.0	0.0±0.0	8.6±0.0	×
BT	65.5±0.2	0.1±0.3	0.0±0.1	×
UNSIR	64.6±0.4	42.9±14.3	40.7±12.1	×
GKT	2.3±0.2	0.0±0.0	56.2±20.0	✓
EMMN	26.9±7.7	24.3±23.7	58.2±14.5	✓
BDSH	66.2±0.1	13.0±0.0	2.9±0.1	✓
OURS	66.2±0.3	14.2±0.6	2.9±0.3	✓

Evaluation

Table 3: (a) VGG-16 Sub-class unlearning performance on CIFAR-20 sub-class Rocket. (b) ViT Sub-class unlearning performance onCIFAR-20 sub-class Rocket

(a)

Method	$\mathcal{D}_r$ Acc.	$\mathcal{D}_f$ Acc.	MIA	ZS
BSLN	75.3±0.0	79.0±0.0	83.1±0.0	×
RTRN	72.9±0.2	11.5±2.8	14.1±1.3	×
FNTN	65.5±0.7	6.2±3.7	22.3±5.5	×
AMN	73.8±0.2	2.4±2.4	3.0±0.9	×
SCRUB	62.4±28.4	10.1±22.48	16.7±21.7	×
SSD	75.0±0.0	4.2±0.0	11.0±0.0	×
BT	74.9±0.2	48.4±16.9	0.1±0.1	×
UNSIR	74.1±0.2	57.5±10.3	57.4±8.6	×
BDSH	74.4±0.0	17.535±0.0	12.9±0.1	✓
OURS	73.7±0.8	19.3±18.3	11.2±7.8	✓

(b)

Method	$\mathcal{D}_r$ Acc.	$\mathcal{D}_f$ Acc.	MIA	ZS
BSLN	95.7 ± 0.0	94.5 ± 0.0	80.4 ± 0.0	×
RTRN	94.6 ± 0.1	22.3 ± 8.3	3.4 ± 1.1	×
FNTN	85.7 ± 3.1	6.2 ± 6.0	16.0 ± 2.7	×
AMN	93.5 ± 0.2	0.8 ± 1.7	0.8 ± 0.3	×
SSD	95.1 ± 0.0	5.12 ± 0.0	5.4 ± 0.0	×
BT	93.6 ± 0.3	3.3 ± 2.9	0.0 ± 0.1	×
UNSIR	93.3 ± 0.4	74.9 ± 10.1	27.3 ± 13.8	×
BDSH	95.7±0.0	48.4±0.0	0.1±0.0	✓
Ours	92.2±0.0	0.0±0.0	14.66±8.8	✓

Table 4: (a) VGG11 Random unlearning performance for 100 samples from CIFAR-10. (b)ViT Random unlearning performance for 100 samples from CIFAR-10.

(a)

Method	$\mathcal{D}_r$ Acc.	$\mathcal{D}_f$ Acc.	MIA	ZS
BSLN	87.0±0.0	92.0±3.6	70.1±5.4	×
RTRN	87.7±0.2	91.0±2.5	78.9±3.5	×
FNTN	84.4±0.8	86.4±4.4	70.8±4.7	×
AMN	86.8±0.3	51.3±4.4	13.1±2.9	×
SCRUB	87.7±0.1	92.7±2.9	71.8±5.2	×
SSD	85.6±2.7	90.8±3.7	66.7±5.9	×
BT	86.9±0.2	82.5±4.9	40.8±6.3	×
BDSH	86.9±0.1	92.2±3.4	69.8±5.1	✓
OURS	86.3±0.3	88.7±3.9	64.2±5.2	✓

(b)

Method	$\mathcal{D}_r$ Acc.	$\mathcal{D}_f$ Acc.	MIA	ZS
BSLN	98.9 ± 0.0	100.0 ± 0.0	90.8 ± 3.5	×
RTRN	98.6 ± 0.1	98.8 ± 0.8	91.8 ± 1.8	×
FNTN	97.3 ± 0.3	97.2 ± 1.0	86.1 ± 2.1	×
AMN	97.6 ± 0.3	73.5 ± 5.1	10.4 ± 4.9	×
SSD	98.0 ± 1.6	98.1 ± 2.4	85.5 ± 0.1	×
BT	97.6 ± 0.4	86.7 ± 3.6	33.5 ± 5.6	×
BDSH	98.0±0.29	97.9±1.6	78.8±0.0	✓
Ours	98.0±0.3	98.0±1.5	78.8±4.0	✓

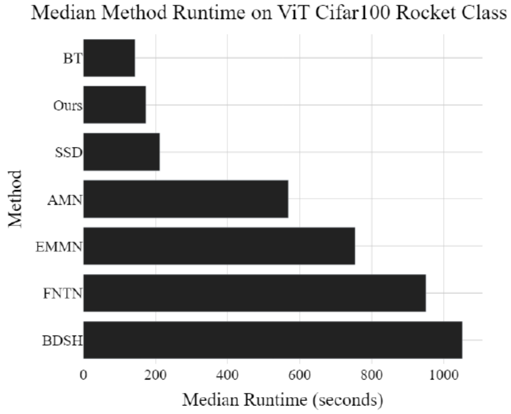


Figure 5: Median method runtime for ViT full-class forgetting on class rocket in seconds. For visual clarity we exclude GKT (∼ 3000 seconds).

Runtime

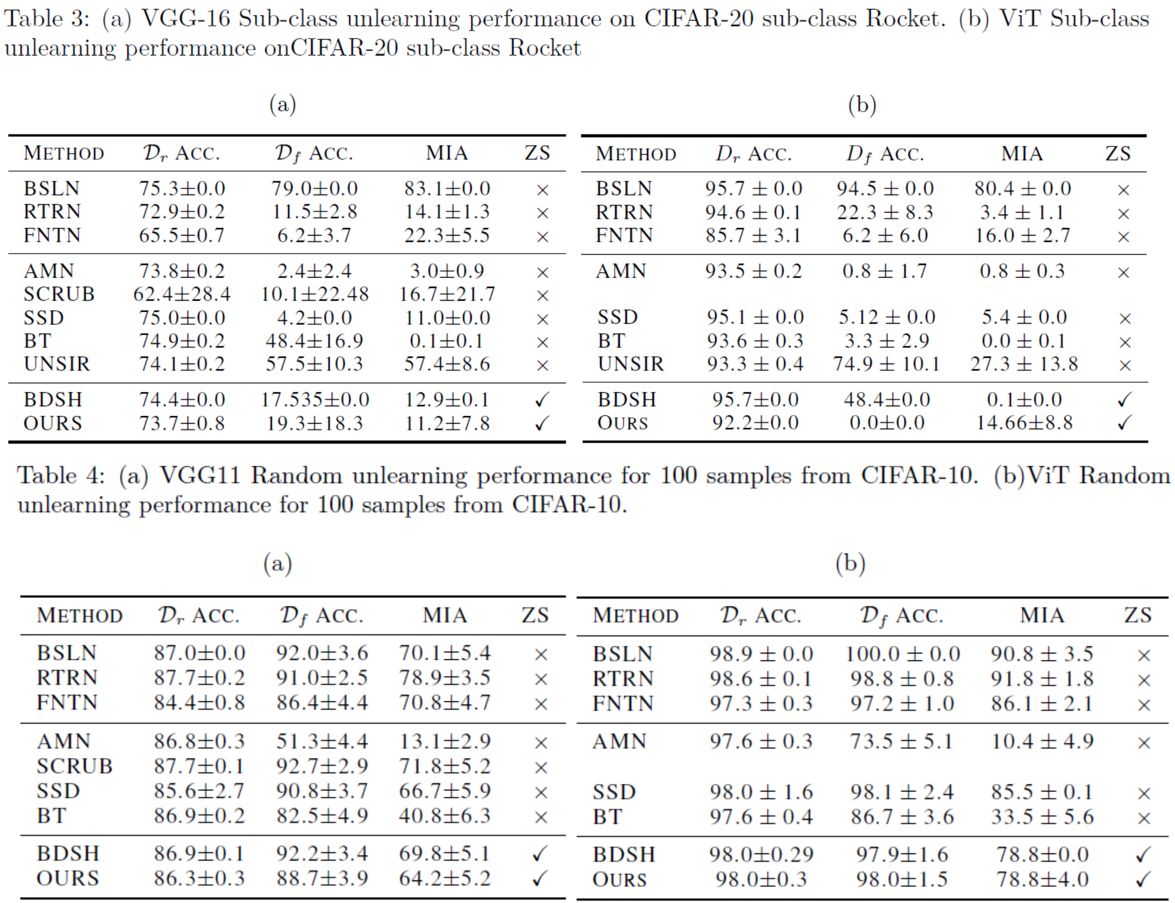


Table 3: (a) VGG-16 Sub-class unlearning performance on CIFAR-20 sub-class Rocket. (b) ViT Sub-class unlearning performance onCIFAR-20 sub-class Rocket

(a)

Method	$\mathcal{D}_r$ Acc.	$\mathcal{D}_f$ Acc.	MIA	ZS
BSLN	75.3±0.0	79.0±0.0	83.1±0.0	×
RTRN	72.9±0.2	11.5±2.8	14.1±1.3	×
FNTN	65.5±0.7	6.2±3.7	22.3±5.5	×
AMN	73.8±0.2	2.4±2.4	3.0±0.9	×
SCRUB	62.4±28.4	10.1±22.48	16.7±21.7	×
SSD	75.0±0.0	4.2±0.0	11.0±0.0	×
BT	74.9±0.2	48.4±16.9	0.1±0.1	×
UNSIR	74.1±0.2	57.5±10.3	57.4±8.6	×
BDSH	74.4±0.0	17.535±0.0	12.9±0.1	✓
OURS	73.7±0.8	19.3±18.3	11.2±7.8	✓

(b)

Method	$\mathcal{D}_r$ Acc.	$\mathcal{D}_f$ Acc.	MIA	ZS
BSLN	95.7 ± 0.0	94.5 ± 0.0	80.4 ± 0.0	×
RTRN	94.6 ± 0.1	22.3 ± 8.3	3.4 ± 1.1	×
FNTN	85.7 ± 3.1	6.2 ± 6.0	16.0 ± 2.7	×
AMN	93.5 ± 0.2	0.8 ± 1.7	0.8 ± 0.3	×
SSD	95.1 ± 0.0	5.12 ± 0.0	5.4 ± 0.0	×
BT	93.6 ± 0.3	3.3 ± 2.9	0.0 ± 0.1	×
UNSIR	93.3 ± 0.4	74.9 ± 10.1	27.3 ± 13.8	×
BDSH	95.7±0.0	48.4±0.0	0.1±0.0	✓
Ours	92.2±0.0	0.0±0.0	14.66±8.8	✓

Table 4: (a) VGG11 Random unlearning performance for 100 samples from CIFAR-10. (b)ViT Random unlearning performance for 100 samples from CIFAR-10.

(a)

Method	$\mathcal{D}_r$ Acc.	$\mathcal{D}_f$ Acc.	MIA	ZS
BSLN	87.0±0.0	92.0±3.6	70.1±5.4	×
RTRN	87.7±0.2	91.0±2.5	78.9±3.5	×
FNTN	84.4±0.8	86.4±4.4	70.8±4.7	×
AMN	86.8±0.3	51.3±4.4	13.1±2.9	×
SCRUB	87.7±0.1	92.7±2.9	71.8±5.2	×
SSD	85.6±2.7	90.8±3.7	66.7±5.9	×
BT	86.9±0.2	82.5±4.9	40.8±6.3	×
BDSH	86.9±0.1	92.2±3.4	69.8±5.1	✓
OURS	86.3±0.3	88.7±3.9	64.2±5.2	✓

(b)

Method	$\mathcal{D}_r$ Acc.	$\mathcal{D}_f$ Acc.	MIA	ZS
BSLN	98.9 ± 0.0	100.0 ± 0.0	90.8 ± 3.5	×
RTRN	98.6 ± 0.1	98.8 ± 0.8	91.8 ± 1.8	×
FNTN	97.3 ± 0.3	97.2 ± 1.0	86.1 ± 2.1	×
AMN	97.6 ± 0.3	73.5 ± 5.1	10.4 ± 4.9	×
SSD	98.0 ± 1.6	98.1 ± 2.4	85.5 ± 0.1	×
BT	97.6 ± 0.4	86.7 ± 3.6	33.5 ± 5.6	×
BDSH	98.0±0.29	97.9±1.6	78.8±0.0	✓
Ours	98.0±0.3	98.0±1.5	78.8±4.0	✓

Discussion, Limitation

- Pros
  - Novel information-theoretic zero-shot unlearning
  - Entropy and accuracy profiles statistically indistinguishable from retraining-from-scratch
  - Consistent performance across three unlearning task as well as low compute overhead

- Cons
  - Require full gradient access
  - Hyperparameter sensitivity
    - Left: lr, center: std for generating noise, right: Drop retain set acc.
  - No mention of exact value of τ used in experiments

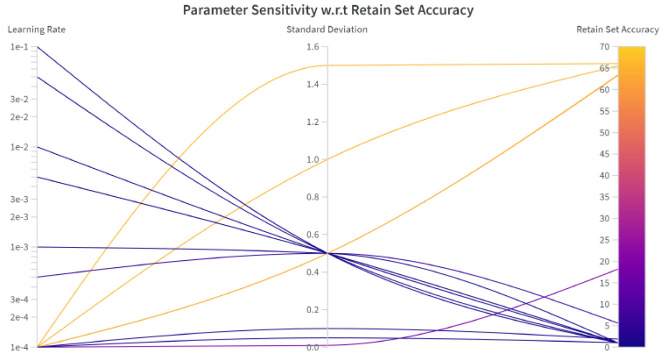
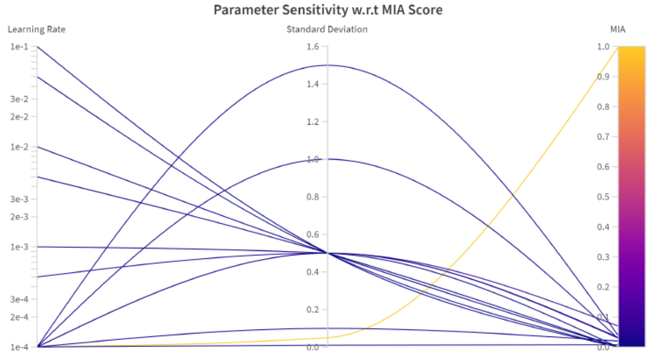


Figure 7: Plot of the $\mathcal{D}_r$ sensitivity to change in hyper-parameters for VGG11 full-class forgetting.



igure 8: Plot of the MIA sensitivity to change in hyper-parameters for VGG11 full-class forgetting.

SecAI Lab

Thank you