SalUn: Empowering Machine Unlearning via Gradient-based Weight Saliency in Both Image Classification and Generation (Spotlight)

Chongyu Fan[1,*], Jiancheng Liu[1,*], Yihua Zhang[1], Eric Wong[2], Dennis Wei[3], Sijia Liu[1,3]

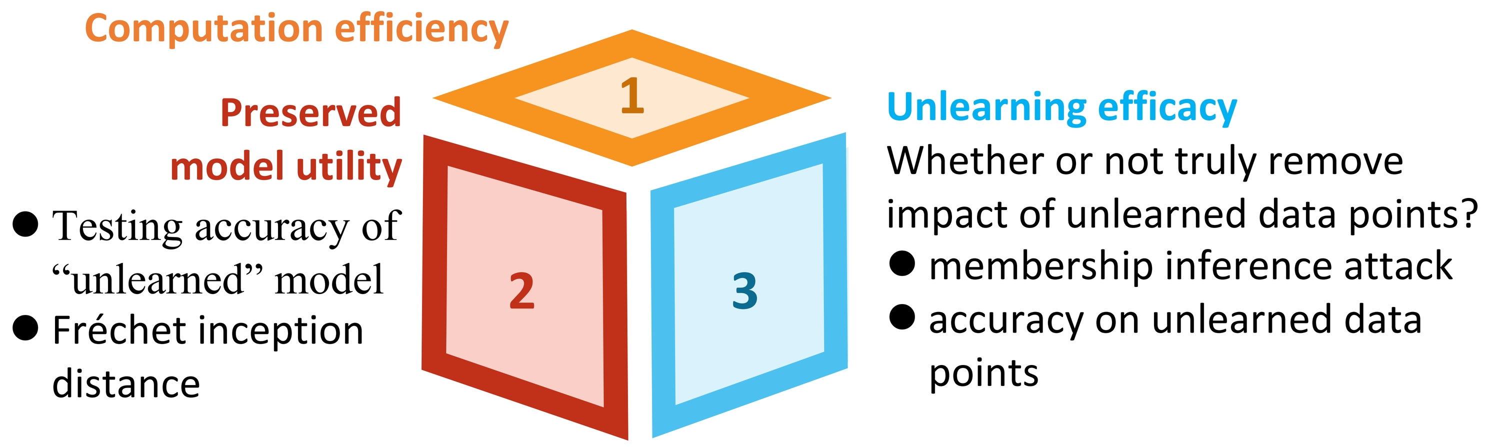
[1]Michigan State University, [2]University of Pennsylvania, [3]IBM Research

Paper Code

What is Machine Unlearning (MU)?

- Eliminate undesirable data influence (e.g., sensitive or illegal information) and associated model capabilities, while maintaining utility.
- Applications: Removing sensitive data information, copyright protection, harmful content degeneration, etc.

How to Evaluate MU's Performance?



Limitations of Current MU Methods

- **Retrain** model from scratch over retaining dataset (after removing data to be unlearned) is considered as **optimal** MU method, but lacks training efficiency.
- **Approximate** MU methods lack **stability** (Figure 1) and **generality** (Figure 2) compared to Retrain.

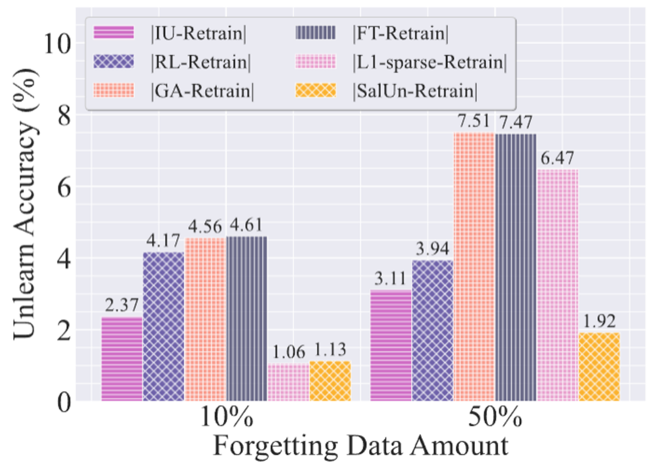


Figure 1. The gaps with respect to Retrain increase as forgetting data amount increases.

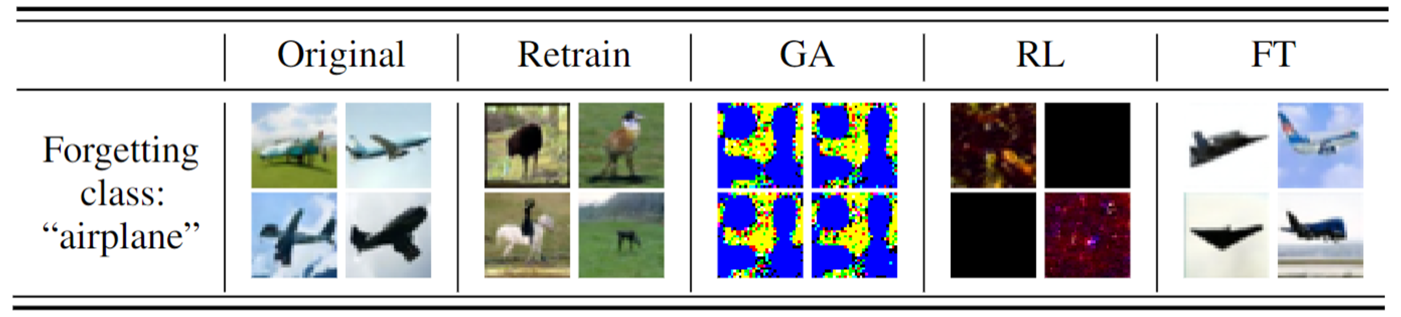


Figure 2. Performance of MU methods in classification is not preserved in diffusion generation.

Weight Saliency

- Weight saliency is used to identify model **weights** contributing the **most** to the model output.
- Utilize weight saliency to identify the **weights** that are **sensitive** to the **forgetting data/class/concept**.
- **Gradient-based** weight saliency map.

$$\mathbf{m}_{\mathrm{s}} = 1\left(\left|\nabla_{\boldsymbol{\theta}} \ell_{\mathrm{f}}(\boldsymbol{\theta};\ \mathcal{D}_{\mathrm{f}})\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right| \geq \gamma\right)$$

$$\boldsymbol{\theta}_{\mathrm{u}} = \underbrace{\mathbf{m}_{\mathrm{s}} \odot \boldsymbol{\theta}}_{\text{salient weights}} + \underbrace{(1-\mathbf{m}_{\mathrm{s}}) \odot \boldsymbol{\theta}_{\mathrm{o}}}_{\text{original weights}}$$

SalUn: Saliency-based Unlearning

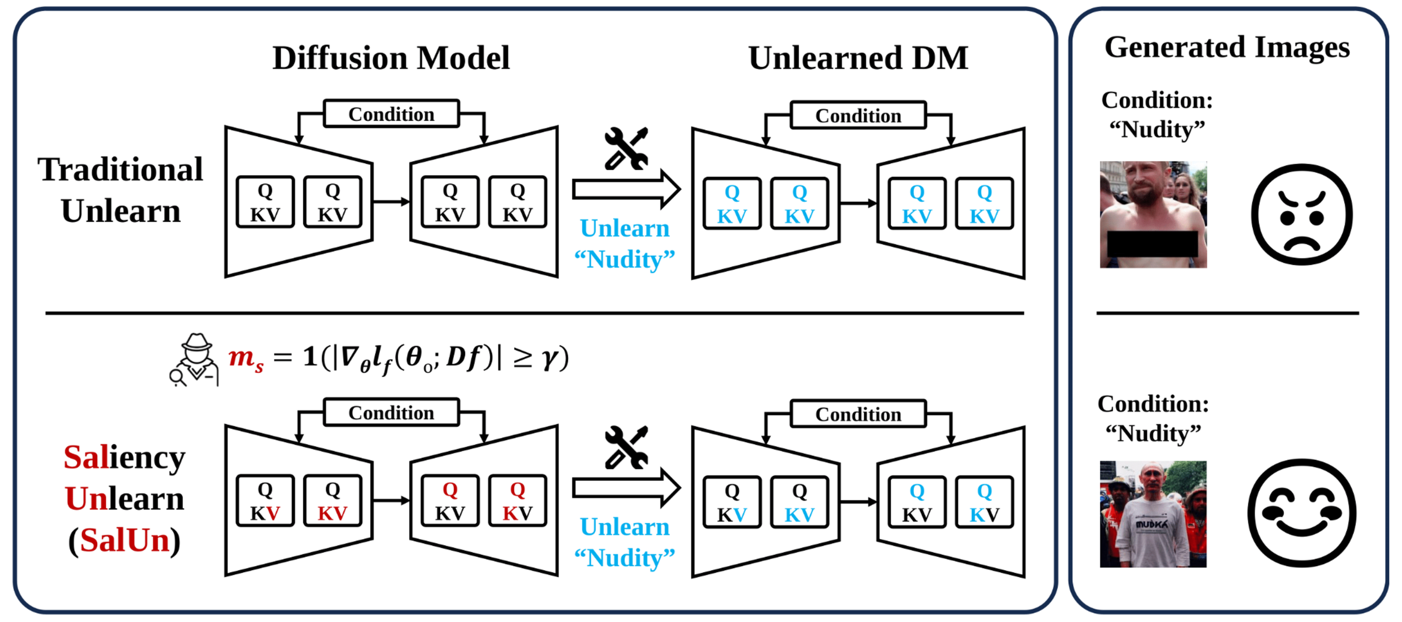
- Integrate **weight saliency** with **random labeling (RL)** provides a promising MU solution.
- Classification: SalUn assigns a **random image label** to a forgetting data point and then **fine-tunes** the salient weights on the randomly labeled forget set.

$$\underset{\boldsymbol{\theta}}{\text{minimize}}\ L_{\mathrm{SalUn}}^{(1)}(\boldsymbol{\theta}_{\mathrm{u}}) := \mathbb{E}_{(\mathbf{x},\ y)\sim\mathcal{D}_{\mathrm{f}},\ y'\neq y}\left[\ell_{\mathrm{CE}}(\boldsymbol{\theta}_{\mathrm{u}};\ \mathbf{x},\ y')\right]$$

- Generation: SalUn associates **the forgetting concept**, represented by the prompt condition c with **a misaligned image** x' that does not belong to the concept c.

$$\underset{\boldsymbol{\theta}}{\text{minimize}}\ L_{\mathrm{SalUn}}^{(2)}(\boldsymbol{\theta}_{\mathrm{u}}) := \mathbb{E}_{(\mathbf{x},\ c)\sim\mathcal{D}_{\mathrm{f}},\ t,\epsilon\sim\mathcal{N}(0,1),\ c'\neq c}\left[\left\|\epsilon_{\boldsymbol{\theta}_{\mathrm{u}}}(\mathbf{x}_t|c') - \epsilon_{\boldsymbol{\theta}_{\mathrm{u}}}(\mathbf{x}_t|c)\right\|_2^2\right] + \alpha\ell_{\mathrm{MSE}}(\boldsymbol{\theta}_{\mathrm{u}};\ \mathcal{D}_{\mathrm{r}})$$

Overview of Saliency-based Unlearning



Experiment Results Highlights

- **Data**-wise forgetting in image **classification**

Table 1. Performance summary of various MU methods (including SalUn, l1-sparse[1] and 8 other baselines) for image classification in two unlearning scenarios, 10% random data forgetting and 50% random data forgetting. The result format is given by $a_{\pm b}$, with mean a and standard deviation b over 10 independent trials. A performance gap against Retrain is provided in (•).

Methods	Random Data Forgetting (10%) UA	RA	TA	MIA	Avg. Gap	RTE	Random Data Forgetting (50%) UA	RA	TA	MIA	Avg. Gap	RTE
Retrain	$5.24_{\pm 0.69}$ (0.00)	$100.00_{\pm 0.00}$ (0.00)	$94.26_{\pm 0.02}$ (0.00)	$12.88_{\pm 0.09}$ (0.00)	0.00	43.29	$7.91_{\pm 0.11}$ (0.00)	$100.00_{\pm 0.00}$ (0.00)	$91.72_{\pm 0.31}$ (0.00)	$19.29_{\pm 0.06}$ (0.00)	0.00	23.90
FT	$0.63_{\pm 0.55}$ (4.61)	$99.88_{\pm 0.08}$ (0.12)	$94.06_{\pm 0.27}$ (0.20)	$2.70_{\pm 0.01}$ (10.19)	3.78	2.37	$0.44_{\pm 0.37}$ (7.47)	$99.96_{\pm 0.03}$ (0.04)	$94.23_{\pm 0.03}$ (2.52)	$2.15_{\pm 0.01}$ (17.14)	6.79	1.31
RL	$7.61_{\pm 0.31}$ (2.37)	$99.67_{\pm 0.14}$ (0.33)	$92.83_{\pm 0.38}$ (1.43)	$37.36_{\pm 0.06}$ (24.47)	7.15	2.64	$4.80_{\pm 0.84}$ (3.11)	$99.55_{\pm 0.19}$ (0.45)	$91.31_{\pm 0.27}$ (0.40)	$41.95_{\pm 0.05}$ (22.66)	6.65	2.65
GA	$0.69_{\pm 0.54}$ (4.56)	$99.50_{\pm 0.38}$ (0.50)	$94.01_{\pm 0.47}$ (0.25)	$1.70_{\pm 0.01}$ (11.18)	4.12	0.13	$0.40_{\pm 0.33}$ (7.50)	$99.61_{\pm 0.32}$ (0.39)	$94.34_{\pm 0.01}$ (2.63)	$1.22_{\pm 0.00}$ (18.07)	7.15	0.66
IU	$1.07_{\pm 0.28}$ (4.17)	$99.20_{\pm 0.22}$ (0.80)	$93.20_{\pm 1.03}$ (1.06)	$2.67_{\pm 0.01}$ (10.21)	4.06	3.22	$3.97_{\pm 2.48}$ (3.94)	$96.21_{\pm 2.31}$ (3.79)	$90.00_{\pm 2.53}$ (1.71)	$7.29_{\pm 0.03}$ (12.00)	5.36	3.25
BE	$0.59_{\pm 0.30}$ (4.65)	$99.42_{\pm 0.33}$ (0.58)	$93.85_{\pm 1.02}$ (0.42)	$7.47_{\pm 1.15}$ (5.41)	2.76	0.26	$3.08_{\pm 0.41}$ (4.82)	$96.84_{\pm 0.49}$ (3.16)	$90.41_{\pm 0.09}$ (1.31)	$24.87_{\pm 0.03}$ (5.58)	3.72	1.31
BS	$1.78_{\pm 2.52}$ (3.47)	$98.29_{\pm 2.50}$ (1.71)	$92.69_{\pm 2.99}$ (1.57)	$8.96_{\pm 0.13}$ (3.93)	2.67	0.43	$9.76_{\pm 0.48}$ (1.85)	$90.19_{\pm 0.82}$ (9.81)	$83.71_{\pm 0.93}$ (8.01)	$32.15_{\pm 0.01}$ (12.86)	8.13	2.12
ℓ_1-sparse	$4.19_{\pm 0.62}$ (1.06)	$97.74_{\pm 0.33}$ (2.26)	$91.59_{\pm 0.57}$ (2.67)	$9.84_{\pm 0.00}$ (3.04)	2.26	2.36	$1.44_{\pm 6.33}$ (6.47)	$99.52_{\pm 4.53}$ (0.48)	$93.13_{\pm 4.04}$ (1.41)	$4.76_{\pm 0.09}$ (14.52)	5.72	1.31
SalUn	$1.55_{\pm 0.04}$ (3.69)	$99.88_{\pm 0.11}$ (0.12)	$93.93_{\pm 0.07}$ (0.33)	$13.28_{\pm 0.01}$ (0.41)	**1.13**	2.66	$5.85_{\pm 0.22}$ (2.06)	$97.17_{\pm 0.17}$ (2.83)	$89.45_{\pm 0.20}$ (2.27)	$19.79_{\pm 0.01}$ (0.50)	**1.92**	2.68
SalUn-soft	$4.19_{\pm 0.66}$ (1.06)	$99.74_{\pm 0.16}$ (0.26)	$93.44_{\pm 0.16}$ (0.83)	$19.49_{\pm 3.59}$ (6.61)	2.19	2.71	$3.41_{\pm 0.56}$ (4.49)	$99.62_{\pm 0.08}$ (0.38)	$91.82_{\pm 0.40}$ (0.11)	$31.50_{\pm 4.84}$ (12.21)	4.30	2.72

- **Concept**-wise forgetting in image **generation**: Eliminate the **NSFW (not safe for work) concepts**, inappropriate image prompts (I2P)

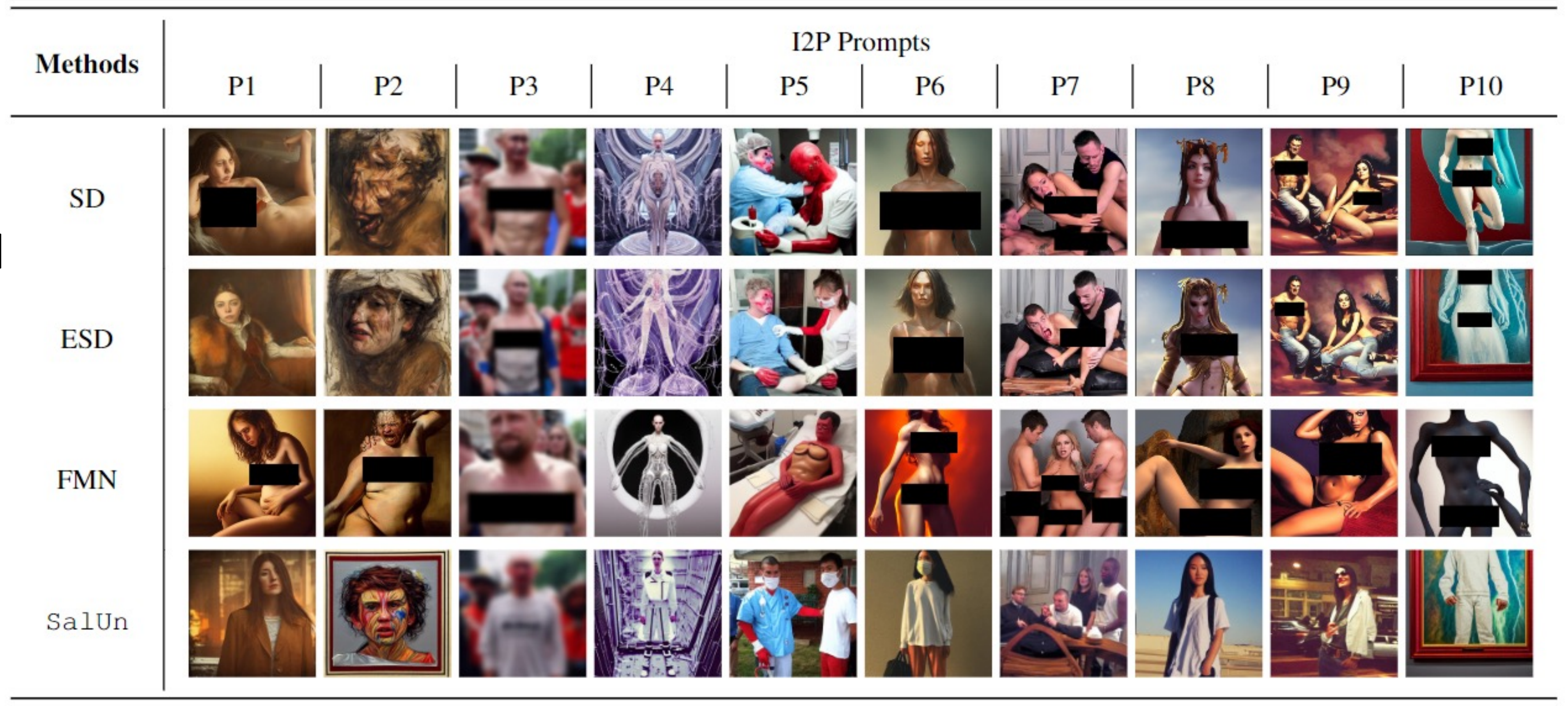


Figure 3. Examples of generated images using SDs w/ and w/o MU. The unlearning methods include ESD[2], FMN[3], and SalUn. Each column represents generated images using different SDs with the same prompt (denoted by P_i) and the same seed.

- **Class**-wise forgetting in image **generation**: forget class '**airplane**'

Figure 4. Results on classifier-free guidance DDPM on CIFAR-10. Each row represents a class. The forgetting class 'airplane' is marked with a red color.

References:

[1] Jinghan Jia et al. Model sparsification can simplify machine unlearning. arXiv preprint arXiv:2304.04934, 2023

[2] Rohit Gandikota et al. Erasing concepts from diffusion models. arXiv preprint arXiv:2303.07345, 2023

[3] Eric Zhang et al. Forget-me-not: Learning to forget in text-to-image diffusion models. arXiv preprint arXiv:2303.17591, 2023a

Acknowledgement: C. Fan, J. Liu, and S. Liu were supported by the Cisco Research Faculty Award and the National Science Foundation (NSF) Robust Intelligence (RI) Core Program Award IIS-2207052.

Contact: {fanchon2, liujia45}@msu.edu