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, [2]IBM Research

*Equal contribution

Paper

Code

{fanchong2, liujia45}@msu.edu

MICHIGAN STATE UNIVERSITY

OPTML

What is Machine Unlearning (MU)?

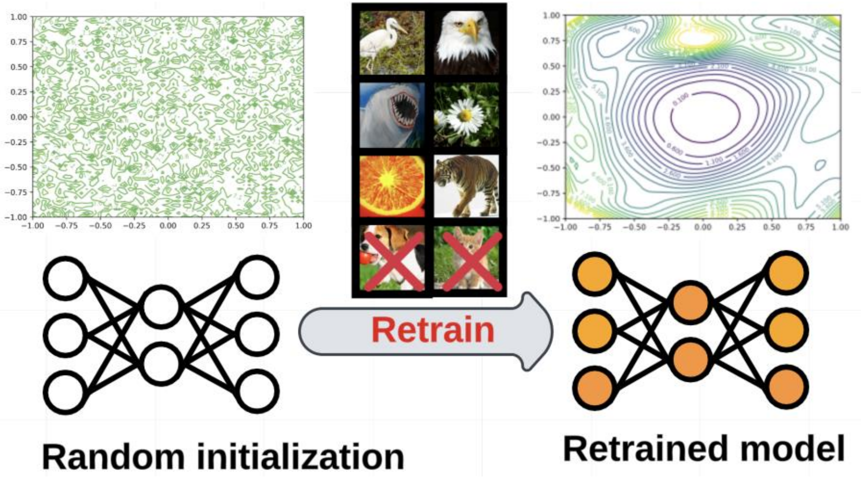
- Eliminate undesirable data influence (e.g., sensitive or illegal information) and associated model capabilities, while maintaining utility.
- Applications: Removing sensitive data information, copyright protection, harmful content degeneration, etc.

Generation examples by Stable Diffusion pre/after applying MU (`SalUn`).

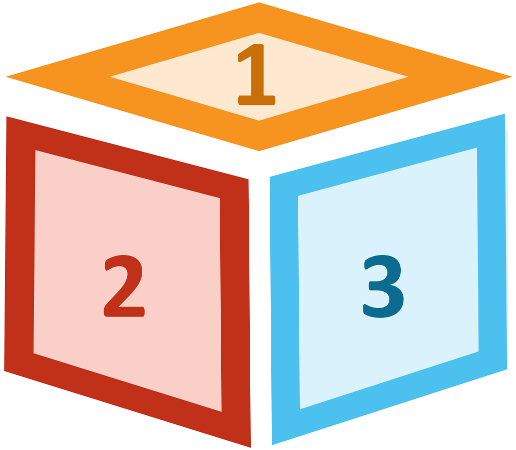
Forgetting Objective: (Left) Concept "*Nudity*"; (Mid) Object "*Dog*"; (Right) Style "*Sketch*".

Why and Why Not Retrain?

- Retrain model from scratch over retaining dataset (after removing data to be unlearned) is considered as an **optimal** MU method.
- Limitation: Lacks training efficiency, particularly for large-scale deep models



How to Define the "GOOD" in MU?



Computation Efficiency

Generalization Fidelity

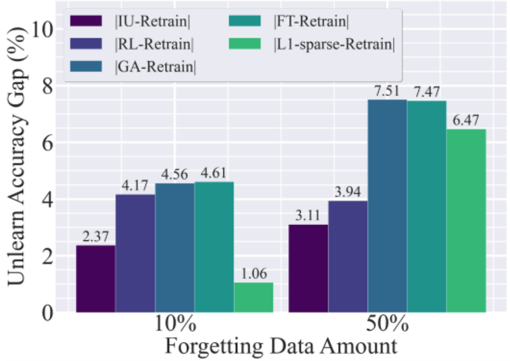
Can unlearned models still generalize?

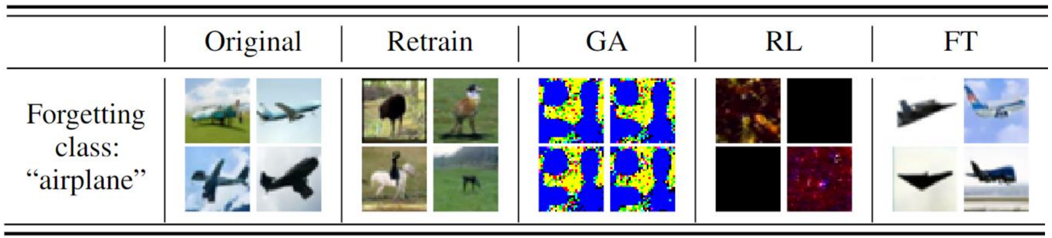
Unlearning Efficacy

Is the impact of forgetting data points truly removed?

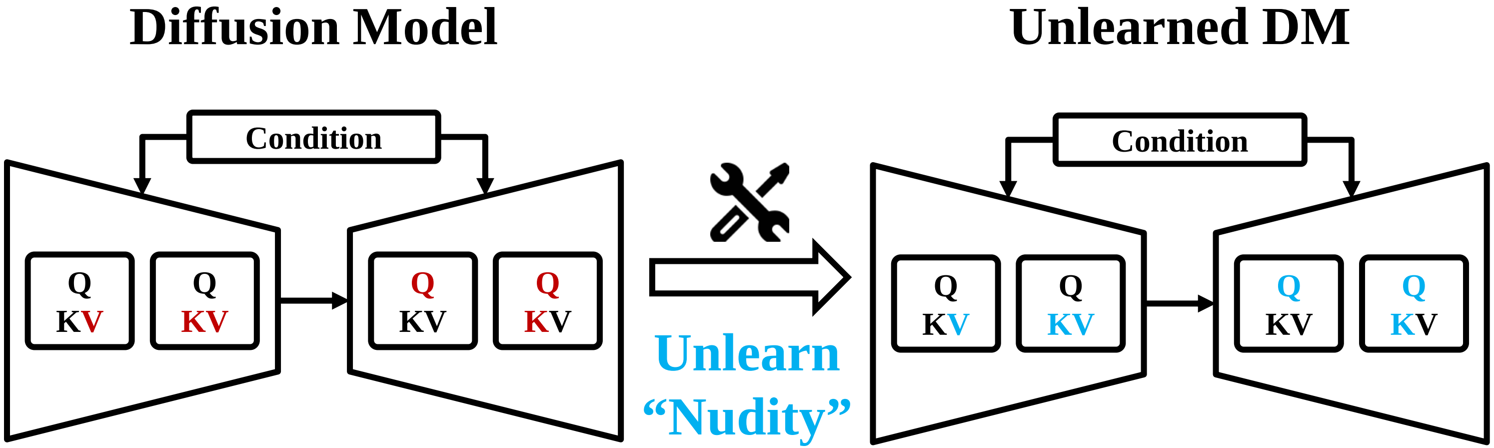
Limitations of Current MU Methods

- Existing MU methods lack generality in harder tasks.
  - Large forgetting ratio
  - Image generation tasks
- Hard to strike the balancing point between unlearning and generalization.
  - Tend to either **over-forget** (e.g., GA , RL)
  - Or **under-forget** (e.g., FT , l1-sparse)





SalUn: Weight Saliency Is the Key to MU



SalUn: Weight Saliency Is the Key to MU

- Use the gradient of the forgetting loss with respect to the model weights $\boldsymbol{\theta}$ under the forgetting dataset.
- Apply a hard thresholding to obtain the weight saliency mask.
- By fixing the low-saliency parameters, achieving an accurate unlearning.

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

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

SalUn: Weight Saliency Is the Key to MU

- Integrate weight saliency with random labeling (RL) provides a promising MU solution both in image classification and image generation.

- Classification:
  - SalUn assigns a random label to each forgetting data point and then fine-tunes the salient weights on the randomly relabeled dataset.

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

SalUn: Weight Saliency Is the Key to MU

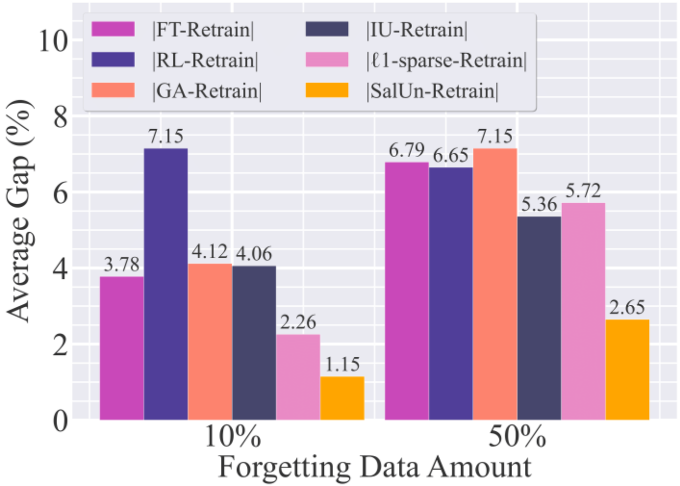
- Integrate weight saliency with random labeling (RL) provides a promising MU solution both in image classification and image generation.

- Generation:
  - SalUn associates each image x in forgetting concept c with a misaligned concept c'.

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

SalUn in Image Classification

- SalUn demonstrates consistent forgetting performance across different forgetting data amounts.



SalUn in Image Generation

- Class-wise forgetting performance in image generation

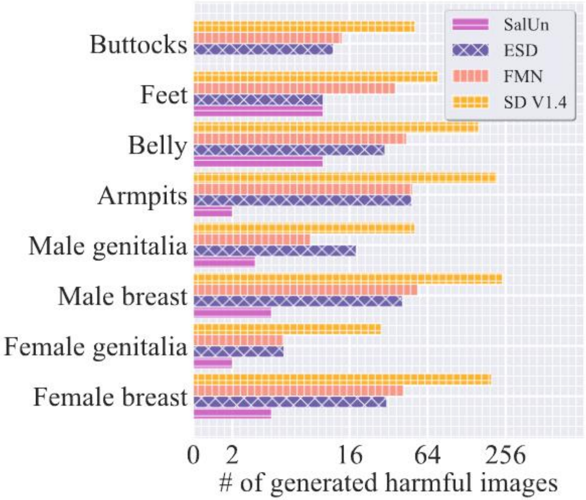
Example of forgetting "*Church*" class.

Forget. Class	SalUn UA (↑)	SalUn FID (↓)	ESD UA (↑)	ESD FID (↓)	FMN UA (↑)	FMN FID (↓)
Tench	**100.00**	2.53	99.40	**1.22**	42.40	1.63
English Springer	**100.00**	**0.79**	100.00	1.02	27.20	1.75
Cassette Player	99.80	0.91	**100.00**	1.84	93.80	**0.80**
Chain Saw	**100.00**	1.58	96.80	1.48	48.40	**0.94**
Church	**99.60**	**0.90**	98.60	1.91	23.80	1.32
French Horn	**100.00**	**0.94**	99.80	1.08	45.00	0.99
Garbage Truck	**100.00**	**0.91**	100.00	2.71	41.40	0.92
Gas Pump	**100.00**	**1.05**	100.00	1.99	53.60	1.30
Golf Ball	98.80	1.45	**99.60**	**0.80**	15.40	1.05
Parachute	**100.00**	1.16	99.80	**0.91**	34.40	2.33
Average	**99.82**	**1.22**	99.40	1.49	42.54	1.30

SalUn in Image Generation

- Concept forgetting performance in image generation

Example of forgetting "*Nudity*" concept.



Summary

- Weight saliency helps Machine Unlearning
- Limitations
  - Limited scope
  - Fine-grained masking methods

terima kasih
mulțumesc
ありがとう
谢谢 ngiyabonga suksema
Met dank
obrigada
Thank
baie dankie
molte grazie
merci
감사합니다
You
Danke schön!
謝謝
obrigado
gracias
Благодарность شكرًا
Спасибі Dziękuję
dank u
mahalo
tusind tak