Not All Wrong is Bad: Using Adversarial Examples for Unlearning

Ali Ebrahimpour-Boroojeny, Hari Sundaram, & Varun Chandrasekaran

University of Illinois at Urbana-Champaign

Introduction

Goal

Removing the influence of data subset $\mathcal{D}_\mathsf{F}$ from a trained model $\mathcal{F}$, so that the resulting model behaves *as if the data were never seen*.

Introduction

Goal

Removing the influence of data subset $\mathcal{D}_\mathsf{F}$ from a trained model $\mathcal{F}$, so that the resulting model behaves *as if the data were never seen*.

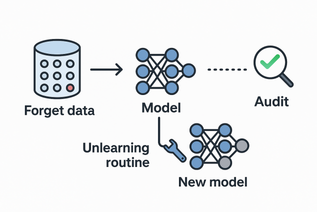
- **Why it matters**
  - Comply with “right to be forgotten” laws (GDPR, CCPA).
  - Remove copyrighted or toxic content in deployed deep learning models.

Introduction

Goal

Removing the influence of data subset $\mathcal{D}_{\mathrm{F}}$ from a trained model $\mathcal{F}$, so that the resulting model behaves *as if the data were never seen*.

- **Why it matters**
  - Comply with "right to be forgotten" laws (GDPR, CCPA).
  - Remove copyrighted or toxic content in deployed deep learning models.
- **Main approaches**
  - *Exact retraining* on $\mathcal{D} - \mathcal{D}_{\mathrm{F}}$
    - ★ Gold standard but costly!
  - *Certified unlearning*
    - ★ Impractical assumptions.
  - *Approximate unlearning*
    - ★ Membership Inference Attacks (MIAs) for evaluations.



Concept: delete subset, apply unlearning routine, audit residual influence.

Basics

- Training set: $\mathcal{D}$
- Forget set: $\mathcal{D}_{\mathsf{F}} \subset \mathcal{D}$
- Remain set: $\mathcal{D}_{\mathsf{R}} = \mathcal{D} - \mathcal{D}_{\mathsf{F}}$

Definition (Machine Unlearning)

Given:

- model architecture $\mathcal{F}$,
- distribution of the learned parameters $\Theta_{\mathcal{D}}$ when $\mathcal{F}$ is trained on $\mathcal{D}$,
- subset $\mathcal{D}_{\mathsf{F}}$ to unlearn,
- distribution of the learned parameters $\Theta_{\mathcal{D}_{\mathsf{F}}}$ when $\mathcal{F}$ is trained on $\mathcal{D}_{\mathsf{R}}$,
- A set of parameters $\boldsymbol{\theta}_{\mathrm{o}} \sim \Theta_{\mathcal{D}}$,

machine unlearning method $\mathcal{M}_{\mathcal{F}}(\theta, \mathcal{D}, \mathcal{D}_{\mathsf{F}})$ gets $\boldsymbol{\theta}_{\mathrm{o}} \sim \Theta_{\mathcal{D}}$ as input and derives a new set of parameters $\boldsymbol{\theta}_{\mathrm{u}} \sim \Theta_{\mathcal{D}_{\mathsf{F}}}$ (aka the unlearned model).

Motivation

Key Observation 1: *The main difference between the predictions on $\mathcal{D}_T$ (unseen samples) and $\mathcal{D}_R$ (observed samples) is that the model's predictions are* much more confident *for the samples that it has observed compared to the unseen samples.*

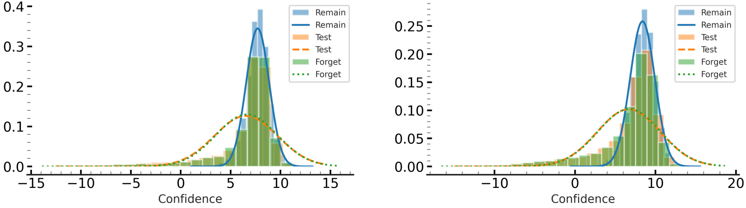
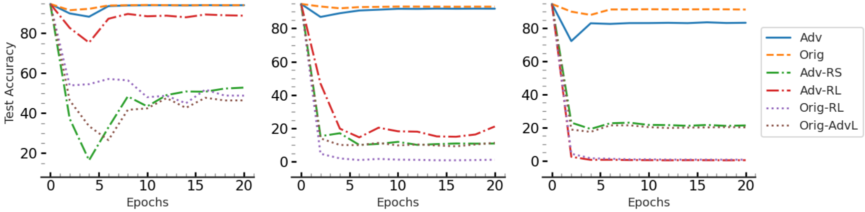


Figure: confidence values of the retrained model for the remaining set (Remain), test set (Test), and forget set (Forget), when the size of the forget set is $\%10$ (1st plot) and $\%50$ (2nd plot) of the training set.

Key Observation 2: *Fine-tuning a model on the adversarial examples does not lead to catastrophic forgetting!*



- ResNet-18 model trained on CIFAR-10
- From left to right, Adv shows fine-tuning on :
  - $\mathcal{D} \cup \mathcal{D}_\mathrm{A}$, $\mathcal{D}_\mathrm{F} \cup \mathcal{D}_\mathrm{A}$, and $\mathcal{D}_\mathrm{A}$

Adversarial set

Algorithm Build Adversarial Set $(\mathcal{F}, \mathcal{A}, \mathcal{D}_{\mathsf{F}}, \epsilon_{init})$

$\mathcal{D}_{\mathsf{A}} = \{\}$
for (x, y) **in** $\mathcal{D}_{\mathsf{F}}$ **do**
    $\epsilon = \epsilon_{init}$
    while TRUE **do**
        $x_{adv} = \mathcal{A}(x, \epsilon)$
        $y_{adv} = \mathcal{F}(x_{adv})$
        if $y_{adv} != y$ **then**
            Break
        end if
        $\epsilon = 2\epsilon$
    end while
    Add (x_{adv}, y_{adv}) to $\mathcal{D}_{\mathsf{A}}$
end for
Return $\mathcal{D}_{\mathsf{A}}$

Evaluation

Unlearning with access to $\mathcal{D}_\mathbf{R}$: Amun outperforms all other methods by achieving lowest Avg. Gap and Amun$_{+SalUn}$ achieves comparable results.

	Random Forget (10%)					**Random Forget** (50%)				
	Unlearn Acc	Retain Acc	Test Acc	FT AUC	Avg. Gap	Unlearn Acc	Retain Acc	Test Acc	FT AUC	Avg. Gap
Retrain	94.49 ± 0.20	100.0 ± 0.00	94.33 ± 0.18	50.00 ± 0.42	0.00	92.09 ± 0.37	100.0 ± 0.00	91.85 ± 0.33	50.01 ± 0.12	0.00
FT	95.16 ± 0.29	96.64 ± 0.25	92.21 ± 0.27	52.08 ± 0.34	2.06 ± 0.10	94.24 ± 0.30	95.82 ± 0.31	91.21 ± 0.33	51.74 ± 0.36	2.17 ± 0.13
RL	95.54 ± 0.14	97.47 ± 0.08	92.17 ± 0.10	51.33 ± 0.63	1.74 ± 0.18	94.83 ± 0.44	99.79 ± 0.04	90.08 ± 0.16	50.78 ± 0.14	1.38 ± 0.09
GA	98.94 ± 1.39	99.22 ± 1.31	93.39 ± 1.18	60.96 ± 2.93	4.28 ± 0.47	100.00 ± 0.00	100.00 ± 0.00	94.65 ± 0.07	63.39 ± 0.26	4.62 ± 0.05
BS	99.14 ± 0.31	99.89 ± 0.06	93.04 ± 0.14	57.85 ± 1.12	3.48 ± 0.32	55.24 ± 5.11	55.67 ± 4.90	50.16 ± 5.28	55.19 ± 0.42	32.01 ± 3.86
l_1-Sparse	94.29 ± 0.34	95.63 ± 0.16	91.55 ± 0.17	51.21 ± 0.32	2.16 ± 0.06	98.00 ± 0.17	98.71 ± 0.13	92.79 ± 0.10	54.44 ± 0.47	2.67 ± 0.11
SalUn	96.25 ± 0.21	98.14 ± 0.16	93.06 ± 0.18	50.88 ± 0.54	1.44 ± 0.12	96.68 ± 0.35	99.89 ± 0.01	91.97 ± 0.18	50.86 ± 0.18	1.36 ± 0.04
Amun	95.45 ± 0.19	99.57 ± 0.00	93.45 ± 0.22	50.18 ± 0.36	**0.62** ± 0.05	93.50 ± 0.09	99.71 ± 0.01	92.39 ± 0.04	49.99 ± 0.18	**0.33** ± 0.03
Amun$_{+SalUn}$	95.02 ± 0.18	99.58 ± 0.04	93.29 ± 0.04	50.72 ± 0.79	<u>0.68</u> ± 0.18	93.56 ± 0.07	99.72 ± 0.02	92.52 ± 0.20	49.81 ± 0.40	<u>0.36</u> ± 0.07

Evaluation (cont.)

Unlearning with access to only $\mathcal{D}_{\mathsf{F}}$: As the results show, $_{+SalUn}$ significantly outperforms all other methods, and achieves comparable results.

	Random Forget (10%)					Random Forget (50%)				
	Unlearn Acc	Retain Acc	Test Acc	FT AUC	Avg. Gap	Unlearn Acc	Retain Acc	Test Acc	FT AUC	Avg. Gap
Retrain	94.49 ±0.20	100.0 ±0.00	94.33 ±0.18	50.00 ±0.42	0.00	92.09 ±0.37	100.0 ±0.00	91.85 ±0.33	50.01 ±0.12	0.00
RL	100.00 ±0.00	100.00 ±0.00	94.45 ±0.09	61.85 ±0.25	4.31 ±0.06	100.00 ±0.00	100.00 ±0.00	94.57 ±0.14	61.99 ±0.10	4.29 ±0.03
GA	4.77 ±3.20	5.07 ±3.54	5.09 ±3.38	49.78 ±0.34	68.53 ±2.45	100.00 ±0.00	100.00 ±0.00	92.65 ±0.09	63.41 ±0.24	5.13 ±0.04
BS	100.00 ±0.00	100.00 ±0.00	94.48 ±0.04	61.41 ±0.29	4.20 ±0.07	100.00 ±0.00	100.00 ±0.00	94.58 ±0.08	62.43 ±0.14	4.40 ±0.05
SalUn	100.00 ±0.00	100.00 ±0.00	94.47 ±0.10	61.09 ±0.40	4.11 ±0.09	100.00 ±0.00	100.00 ±0.00	94.59 ±0.12	62.45 ±0.37	4.40 ±0.07
Amun	94.28 ±0.37	97.47 ±0.10	91.67 ±0.04	52.24 ±0.23	<u>1.94</u> ±0.13	92.77 ±0.52	95.66 ±0.25	89.43 ±0.19	52.60 ±0.22	<u>2.51</u> ±0.09
Amun$_{+SalUn}$	94.19 ±0.38	97.71 ±0.06	91.79 ±0.12	51.93 ±0.12	**1.77** ±0.06	91.90 ±0.63	96.59 ±0.31	89.98 ±0.44	52.32 ±0.56	**2.00** ±0.17

Theoretical results

Recall

AMUN gets $\boldsymbol{\theta}_{\mathrm{o}} \sim \Theta_{\mathcal{D}}$ as input and derives a new set of parameters θ'. The set of parameters $\boldsymbol{\theta}_{\mathrm{u}} \sim \Theta_{\mathcal{D}_{\mathsf{F}}}$ is derived when retraining the model from scratch on $\mathcal{D}_{\mathsf{R}}$.

Theoretical results

Recall

AMUN gets $\boldsymbol{\theta}_{\mathrm{o}} \sim \Theta_{\mathcal{D}}$ as input and derives a new set of parameters θ'. The set of parameters $\boldsymbol{\theta}_{\mathrm{u}} \sim \Theta_{\mathcal{D}_\mathsf{F}}$ is derived when retraining the model from scratch on $\mathcal{D}_\mathsf{R}$.

- We derive an upper-bound on $\|\theta' - \theta_u\|_2$.
    - used as a proxy for the difference from the retrained model.

Theoretical results

Recall

AMUN gets $\boldsymbol{\theta}_{\mathrm{o}} \sim \Theta_{\mathcal{D}}$ as input and derives a new set of parameters θ'. The set of parameters $\boldsymbol{\theta}_{\mathrm{u}} \sim \Theta_{\mathcal{D}_{\mathsf{F}}}$ is derived when retraining the model from scratch on $\mathcal{D}_{\mathsf{R}}$.

- We derive an upper-bound on $\|\theta' - \theta_u\|_2$.
  - used as a proxy for the difference from the retrained model.
- The implications of the theoretical results justifies the design choices in AMUN and instructs how to improve the results.

The following factors enhances the quality of unlearning with AMUN:

Theoretical results (implications)

The following factors enhances the quality of unlearning with AMUN:

- Adversarial examples that are closer to the original samples.

Theoretical results (implications)

The following factors enhances the quality of unlearning with AMUN:

- Adversarial examples that are closer to the original samples.
- Higher quality of adversarial example.

Theoretical results (implications)

The following factors enhances the quality of unlearning with AMUN:

- Adversarial examples that are closer to the original samples.
- Higher quality of adversarial example.
- Transferability of the adversarial example generated on the original model to the retrained model.

Theoretical results (implications)

The following factors enhances the quality of unlearning with AMUN:

- Adversarial examples that are closer to the original samples.
- Higher quality of adversarial example.
- Transferability of the adversarial example generated on the original model to the retrained model.
- Preventing from overfitting to the adversarial example.

Theoretical results (implications)

The following factors enhances the quality of unlearning with AMUN:

- Adversarial examples that are closer to the original samples.
- Higher quality of adversarial example.
- Transferability of the adversarial example generated on the original model to the retrained model.
- Preventing from overfitting to the adversarial example.
- The generalization of the retrained model to the unseen samples.
    - Implying better results when the forget set is smaller.

Theoretical results (implications)

The following factors enhances the quality of unlearning with AMUN:

- Adversarial examples that are closer to the original samples.
- Higher quality of adversarial example.
- Transferability of the adversarial example generated on the original model to the retrained model.
- Preventing from overfitting to the adversarial example.
- The generalization of the retrained model to the unseen samples.
  - Implying better results when the forget set is smaller.
- A lower Lipschitz constant of the model.

Thank You!