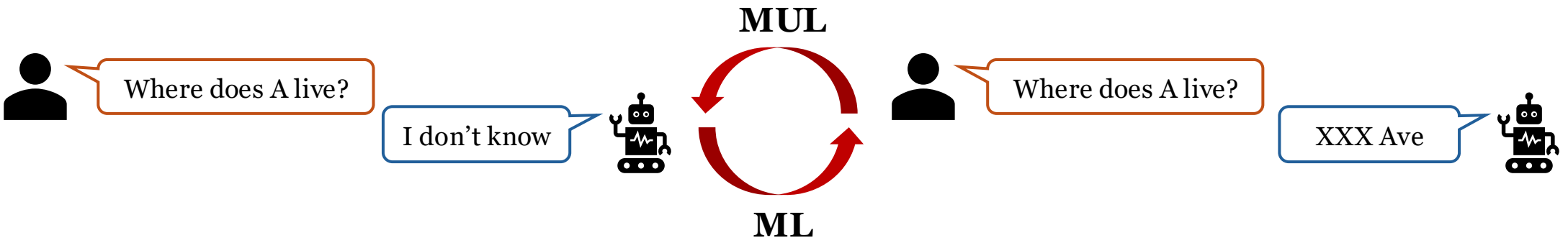
Trustworthy Machine Unlearning: Certification and Verification

Jundong Li, Associate Professor
University of Virginia

https://jundongli.github.io/
jundong@virginia.edu

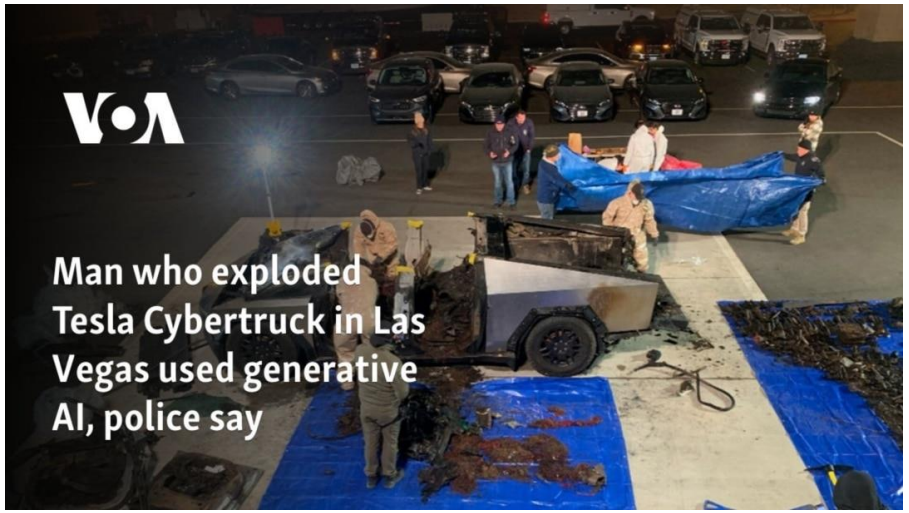
What is Machine Unlearning

- The *right to be forgotten*: data owners have the right to delete their data from an entity storing it
- Data removal in databases ⟶ Data removal in AI models
- Machine Learning v.s. Machine Unlearning



What is Machine Unlearning

- Applications



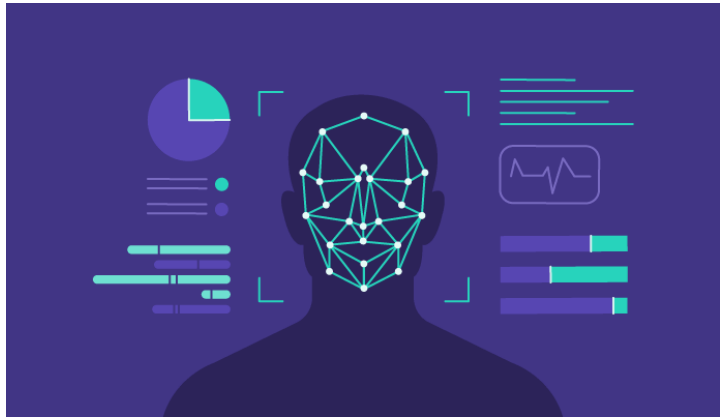
Harmful Knowledge

Copyrighted Data

OpenAI

Vs

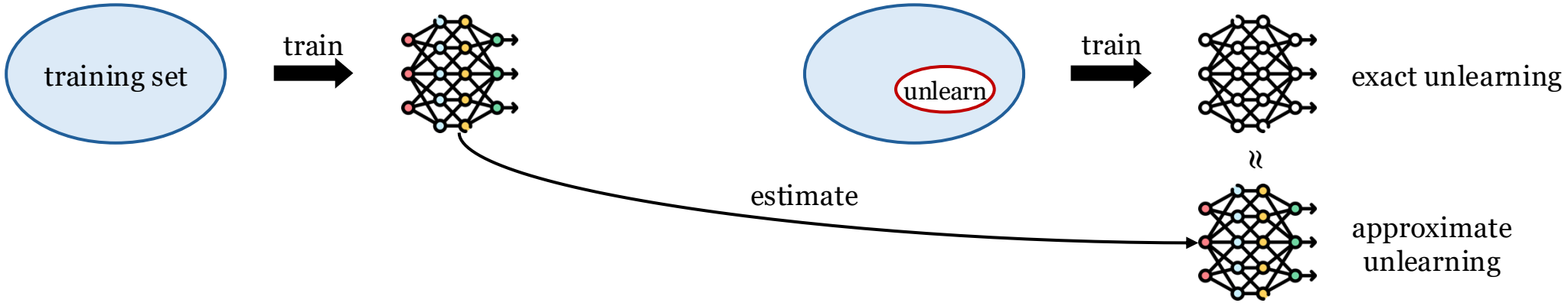
The New York Times



Privacy Leakage

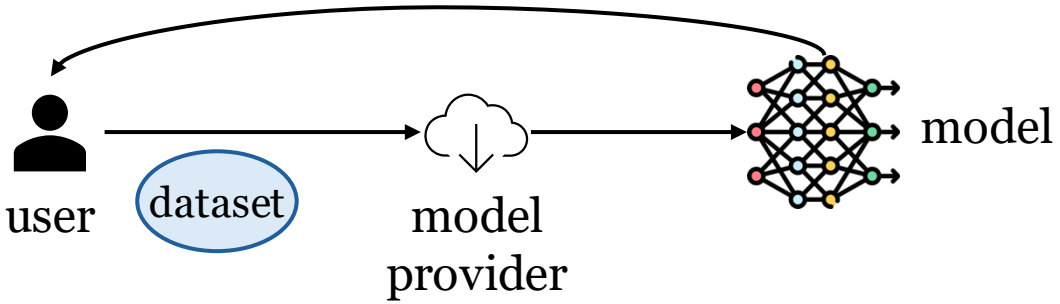
What is Machine Unlearning

- A gold standard: retraining from scratch
- Practical strategies: **exact** and **approximate** unlearning
- Exact unlearning: retraining from scratch more efficiently (sharding)
- Approximate unlearning: estimate retrained model using current model



Trustworthy Machine Unlearning

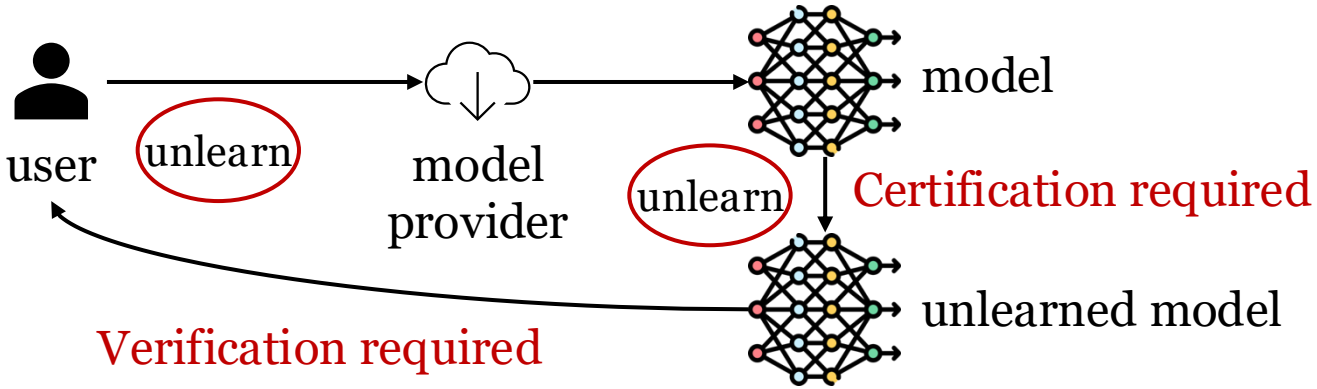
- Two challenges towards trustworthy MUL
  - Certification: how do I **guarantee** my unlearning is good enough
  - Verification: how do **others** know my unlearning is effective



Machine Learning as a Service

Trustworthy Machine Unlearning

- Two challenges towards trustworthy MUL
  - Certification: how do I **guarantee** my unlearning is good enough
  - Verification: how do **others** know my unlearning is effective



Trustworthy Machine Unlearning

Certified Unlearning

- Intuition: Bound the similarity between retrained and unlearned models
- Differential Privacy & Certified Unlearning:

DP

Definition 1. A randomized mechanism $\mathcal{M}: \mathcal{D} \to \mathcal{R}$ with domain $\mathcal{D}$ and range $\mathcal{R}$ satisfies (ε, δ)-differential privacy if for any two adjacent inputs $d, d' \in \mathcal{D}$ and for any subset of outputs $S \subseteq \mathcal{R}$ it holds that

$$\Pr[\mathcal{M}(d) \in S] \leq e^{\varepsilon} \Pr[\mathcal{M}(d') \in S] + \delta.$$

Certified Unlearning

Definition 2 ((ε, δ)-unlearning)**.** *For all S of size n and delete requests $U \subseteq S$ such that $|U| \leq m$, and $W \subseteq \mathcal{W}$, a learning algorithm A and an unlearning algorithm $\bar{A}$ is (ε, δ)-unlearning if*

$$\Pr\big(\bar{A}(U, A(S), T(S)) \in W\big) \leq e^{\varepsilon} \cdot \Pr\big(\bar{A}(\emptyset, A(S \setminus U), T(S \setminus U)) \in W\big) + \delta,$$

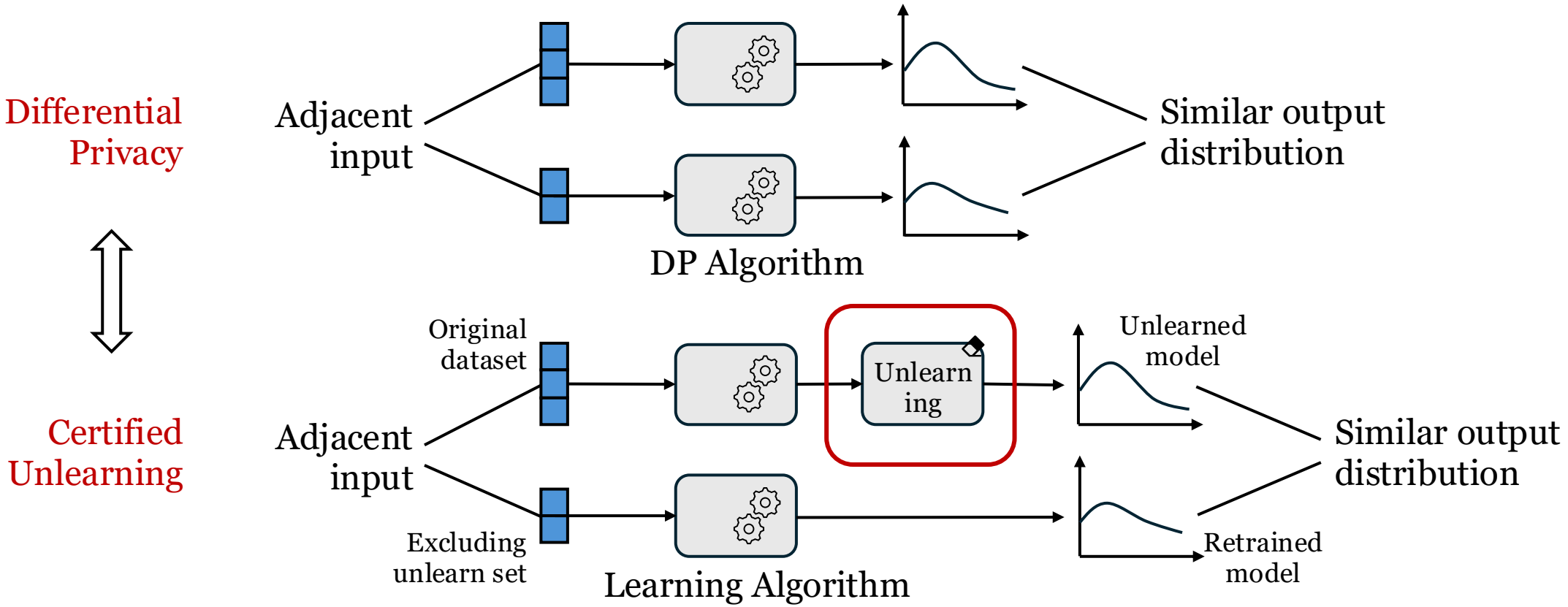
and

$$\Pr\big(\bar{A}(\emptyset, A(S \setminus U), T(S \setminus U)) \in W\big) \leq e^{\varepsilon} \cdot \Pr\big(\bar{A}(U, A(S), T(S)) \in W\big) + \delta,$$

where $\emptyset$ denotes the empty set and $T(S)$ denotes the data statistics available to $\bar{A}$.

Certified Unlearning

- Intuition: Bound the similarity between retrained and unlearned models
- Differential Privacy & Certified Unlearning:



Certified Unlearning

- (ε, δ)-unlearning = Estimate (Newton's/GD) + Gaussian mechanism (noise)
- Certification level (ε, δ) depends on **estimation error**
- Challenge: bounding estimation error for non-convex deep neural nets
- Solution:

Lemma 3.3. *Let* $\boldsymbol{w}^* = \text{argmin}_{\boldsymbol{w}\in\mathcal{H}}\mathcal{L}(\boldsymbol{w}, \mathcal{D})$ *and* $\tilde{\boldsymbol{w}}^* = \text{argmin}_{\boldsymbol{w}\in\mathcal{H}}\mathcal{L}(\boldsymbol{w}, \mathcal{D}_r)$. *Let* $\tilde{\boldsymbol{w}} = \boldsymbol{w}^* - \boldsymbol{H}_{\boldsymbol{w}^*}^{-1}\nabla\mathcal{L}(\boldsymbol{w}^*, \mathcal{D}_r)$ *be an approximation of* $\tilde{\boldsymbol{w}}^*$. *Consider Assumption 3.2, we have*

$$\|\tilde{\boldsymbol{w}} - \tilde{\boldsymbol{w}}^*\|_2 \leq \frac{M}{2}\|\boldsymbol{H}_{\boldsymbol{w}^*}^{-1}\|_2 \cdot \|\boldsymbol{w}^* - \tilde{\boldsymbol{w}}^*\|_2^2. \quad (6)$$

Local convex approximation (adding l-2 reg)

Projected gradient descent (PGD)

"Towards Certified Unlearning for Deep Neural Networks", Binchi Zhang, Yushun Dong, Tianhao Wang, Jundong Li, International Conference on Machine Learning (ICML), 2024.

Certified Unlearning

- Results

Table 2. Comparison between the certified unlearning method and unlearning baselines over three popular DNNs across three real-world datasets. We record the relearn time, the accuracy of the membership inference attack, and the AUC score of the membership inference attack for measuring the unlearning performance.

Method	MLP & MNIST			AllCNN & CIFAR-10			ResNet18 & SVHN		
	Relearn T	Attack Acc	Attack AUC	Relearn T	Attack Acc	Attack AUC	Relearn T	Attack Acc	Attack AUC
Retrain	25	93.10 ± 0.33	95.16 ± 0.47	17	79.82 ± 0.35	88.71 ± 0.43	7	90.47 ± 0.14	93.07 ± 0.27
Fine Tune	17	93.65 ± 0.23	95.37 ± 0.46	14	79.42 ± 1.05	88.13 ± 0.66	7	90.63 ± 0.32	92.96 ± 0.31
Neg Grad	21	93.73 ± 0.45	95.42 ± 0.43	17	78.63 ± 1.23	87.58 ± 0.96	9	90.02 ± 0.13	92.89 ± 0.22
Fisher	21	93.85 ± 0.22	95.37 ± 0.51	14	79.70 ± 1.03	88.58 ± 0.76	9	90.47 ± 0.84	93.13 ± 0.19
L-CODEC	20	95.05 ± 0.05	95.31 ± 0.21	14	83.60 ± 0.62	92.18 ± 0.17	7	93.22 ± 0.35	93.75 ± 0.54
Certified	24	93.22 ± 0.46	95.28 ± 0.50	25	78.00 ± 1.18	87.22 ± 1.13	9	88.63 ± 1.58	92.18 ± 1.16

- Evaluation metrics: relearn time, membership inference attack acc & AUC
- Our method is more effective than other unlearning baselines in removing the information of the unlearned samples

Certified Unlearning

- Results

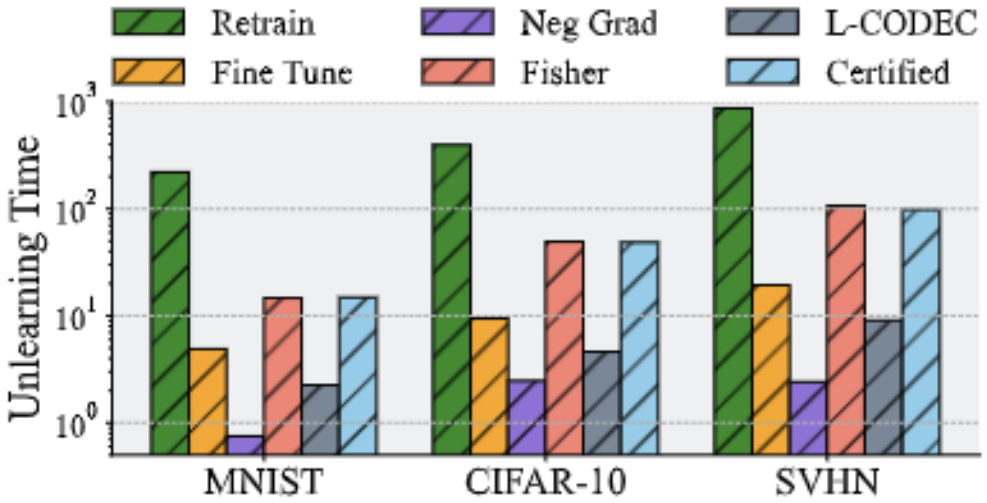


Figure 2. Comparison of unlearning time between the certified unlearning method and unlearning baselines over three popular DNNs across three datasets.

- Certified unlearning has over 10 times speedup compared with exact unlearning (retrain) over DNN

Unlearning Verification

"Verification of Machine Unlearning is Fragile", Binchi Zhang, Zihan Chen, Cong Shen, Jundong Li, International Conference on Machine Learning (ICML), 2024.

- Unlearning protects user privacy, but does not benefit model provider
- Threat Model (Dishonest Model Provider):
  - Adversary's Goal: Pretend to unlearn but actually not (better model utility and lower computational cost)
  - Adversary's Knowledge: complete access to the training process, e.g., training data, unlearned data, and model
- Verification of Machine Unlearning
  - Backdoor: Label some dog images as "cat". If unlearned, they will be correctly recognized as dog
  - Reproduce: Provide a reproducible **proof of retraining** for a verifier. If unlearned, **reproducible** and **unlearned data excluded**

Current verification strategies are fragile

Unlearning Verification

- Adversarial Unlearning

Algorithm 1 Retraining-based Adversarial Unlearning Algorithm

Input: Training data $\mathcal{D}$, unlearned data $\mathcal{D}_u$.
Output: Proof of Retraining $\mathcal{P}_r$.
Initialize $\boldsymbol{w}_r^{(0)}$ and $\mathcal{P}_r \leftarrow \emptyset$.
for $t = 1$ **to** T **do**
  Uniform mini-batch sampling $d^{(t)} \in \mathcal{D}$.
  Choose $d_r^{(t)} \leftarrow \mathcal{S}_r(\boldsymbol{w}_r^{(t-1)}; d^{(t)})$ or $d_r^{(t)} \leftarrow \mathcal{S}_n(d^{(t)})$.
  $\boldsymbol{w}_r^{(t)} \leftarrow g_r^{(t)}(\boldsymbol{w}_r^{(t-1)}, d_r^{(t)})$.
  $\mathcal{P}_r \leftarrow \mathcal{P}_r \cup (\boldsymbol{w}_r^{(t)}, d_r^{(t)}, g_r^{(t)})$.
end for

Replacing the unlearned data in the mini-batch with its nearest neighbor in the remaining dataset

Algorithm 2 Forging-based Adversarial Unlearning Algorithm

Input: Training data $\mathcal{D}$, unlearned data $\mathcal{D}_u$, closest-neighbor mapping $\mathcal{N} : \mathcal{D}_u \rightarrow \mathcal{D} \backslash \mathcal{D}_u$, Proof of Training $\mathcal{P}_t = \{\boldsymbol{w}^{(t)}, d^{(t)}, g^{(t)}\}$.
Output: Proof of Retraining $\mathcal{P}_r$.
$\mathcal{P}_r \leftarrow \emptyset$.
for $t = 1$ **to** T **in parallel do**
  if $d^{(t)} \cap \mathcal{D}_u = \emptyset$ **then**
    $d_r^{(t)} \leftarrow d^{(t)}$.
    $\boldsymbol{w}_r^{(t)} \leftarrow \boldsymbol{w}^{(t)} - \gamma_r^{(t)} \nabla l(f_{\boldsymbol{w}^{(t)}}(\boldsymbol{x}^{(t)}), y^{(t)})$.
  else
    $d_r^{(t)} \leftarrow \mathcal{S}_n(d^{(t)})$.
    $\boldsymbol{w}_r^{(t)} \leftarrow g^{(t)}(\boldsymbol{w}^{(t-1)}, d_r^{(t)})$.
  end if
  $g_r^{(t)} \leftarrow g^{(t)}$.
  $\mathcal{P}_r \leftarrow \mathcal{P}_r \cup (\boldsymbol{w}_r^{(t)}, d_r^{(t)}, g_r^{(t)})$.
end for

Forge the proof of retraining by replacing the unlearned data with nearest neighbor in the remaining dataset

Unlearning Verification

- Adversarial Unlearning

Algorithm 1 Retraining-based Adversarial Unlearning Algorithm

Input: Training data $\mathcal{D}$, unlearned data $\mathcal{D}_u$.
Output: Proof of Retraining $\mathcal{P}_r$.
Initialize $\boldsymbol{w}_r^{(0)}$ and $\mathcal{P}_r \leftarrow \emptyset$.
for $t = 1$ **to** T **do**
  Uniform mini-batch sampling $d^{(t)} \in \mathcal{D}$.
  Choose $d_r^{(t)} \leftarrow \mathcal{S}_r(\boldsymbol{w}_r^{(t-1)}; d^{(t)})$ or $d_r^{(t)} \leftarrow \mathcal{S}_n(d^{(t)})$.
  $\boldsymbol{w}_r^{(t)} \leftarrow g_r^{(t)}(\boldsymbol{w}_r^{(t-1)}, d_r^{(t)})$.
  $\mathcal{P}_r \leftarrow \mathcal{P}_r \cup (\boldsymbol{w}_r^{(t)}, d_r^{(t)}, g_r^{(t)})$.
end for

Replacing the unlearned data in the mini-batch with its nearest neighbor in the remaining dataset

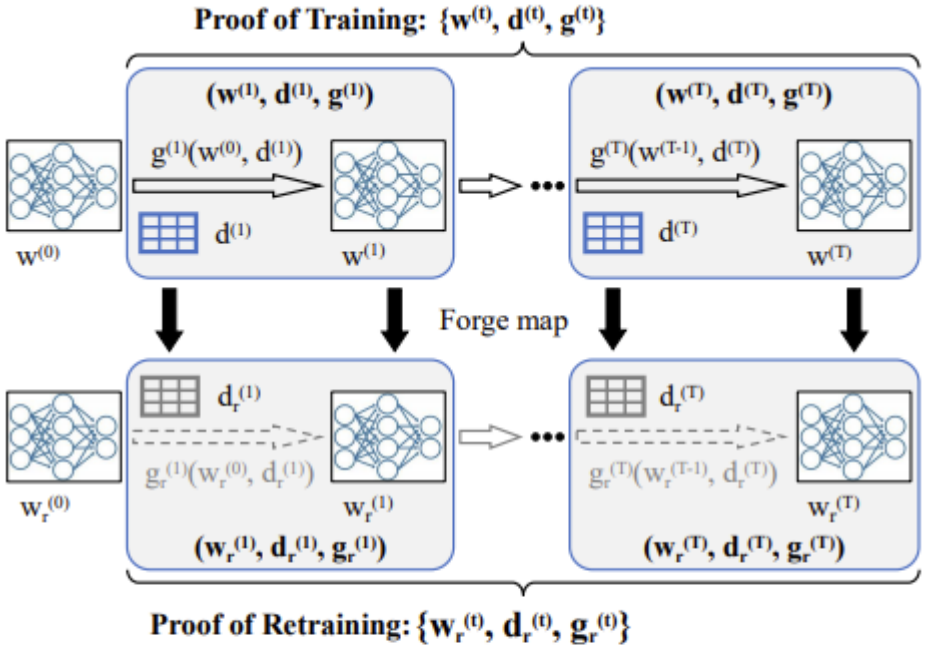


Figure 3. An illustration of the forging-based adversarial unlearning framework. Different from the retraining-based adversarial method, the PoRT here is generated directly from the PoT recorded in the original training. $\boldsymbol{w}_r^{(t)}$ (with $d_r^{(t)}$) is obtained by conducting the forging map over the PoT instead of using the model updating function $g_r^{(t)}$.

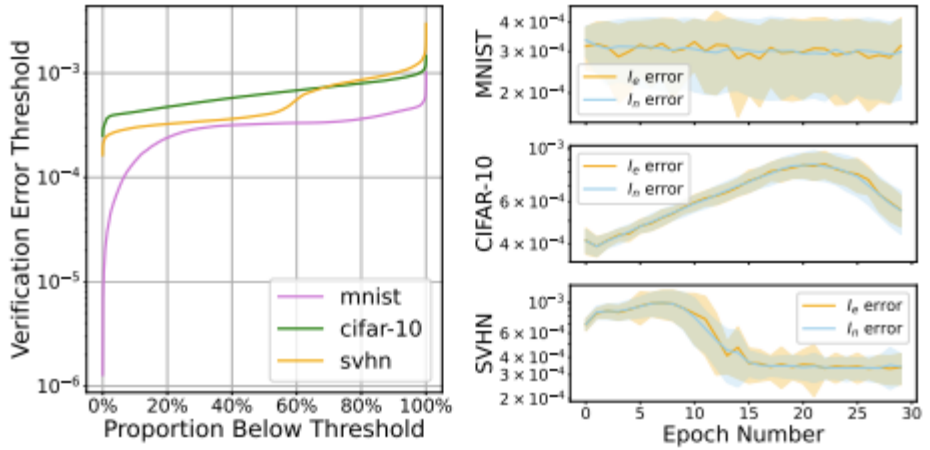
the proof of
ning by
cing the
rned data
nearest
bor in the
ining dataset

Unlearning Verification

- Results

Table 2. Probability of the type II error (β value) of backdoor verification on different (adversarial) unlearning strategies over three datasets.

Method	MNIST	CIFAR-10	SVHN
Original	2.61×10^{-42}	5.14×10^{-20}	1.11×10^{-28}
Retrain	0.998	0.998	0.999
Adv-R	0.997	0.997	0.999
Adv-F	5.46×10^{-34}	2.78×10^{-16}	2.08×10^{-26}



(a) Verification error (b) Details of error statistics

Figure 4. Verification error of forging-based adversarial unlearning method for MLP over MNIST, CNN over CIFAR-10, and ResNet over SVHN.

1. Alg 1 and naive retraining are almost always regarded as truly unlearning the target data
2. Alg 2 and original training are hardly regarded as truly unlearning the target data

1. Alg 2 can pass the reproducing verification within verification error threshold of 0.001
2. The verification error of a checkpoint is determined by the gradient norm of the checkpoint (under fixed learning rate)

Unlearning Verification

- Results

Table 3. Comparison of the model utility among original training, naive retraining, and adversarial unlearning methods over three popular DNNs across three real-world datasets. We record the macro F1-score of the predictions on the unlearned set $\mathcal{D}_u$, retained set $\mathcal{D}\backslash\mathcal{D}_u$, and test set $\mathcal{D}_t$. The prefix 'im-' denotes the results in the class-imbalanced setting.

Method	MLP & MNIST			CNN & CIFAR-10			ResNet & SVHN		
	$\mathcal{D}_u$	$\mathcal{D}\backslash\mathcal{D}_u$	$\mathcal{D}_t$	$\mathcal{D}_u$	$\mathcal{D}\backslash\mathcal{D}_u$	$\mathcal{D}_t$	$\mathcal{D}_u$	$\mathcal{D}\backslash\mathcal{D}_u$	$\mathcal{D}_t$
Original	99.47 ± 0.09	99.76 ± 0.08	97.00 ± 0.17	100.00 ± 0.00	100.00 ± 0.00	85.33 ± 0.31	100.00 ± 0.00	100.00 ± 0.00	94.91 ± 0.09
Retrain	96.43 ± 0.19	99.52 ± 0.10	96.75 ± 0.13	83.60 ± 0.31	100.00 ± 0.00	83.12 ± 0.23	94.33 ± 0.24	100.00 ± 0.00	94.57 ± 0.06
Adv-R ($\mathcal{S}_r$)	98.17 ± 0.16	99.33 ± 0.18	96.78 ± 0.13	83.81 ± 0.44	100.00 ± 0.00	83.08 ± 0.34	94.38 ± 0.11	100.00 ± 0.00	94.54 ± 0.09
Adv-R ($\mathcal{S}_n$)	96.34 ± 0.11	98.65 ± 0.19	96.60 ± 0.14	82.40 ± 0.39	100.00 ± 0.00	81.85 ± 0.44	94.64 ± 0.20	100.00 ± 0.00	94.75 ± 0.04
Adv-F	99.30 ± 0.13	99.33 ± 0.10	96.94 ± 0.14	100.00 ± 0.00	100.00 ± 0.00	85.20 ± 0.24	100.00 ± 0.00	100.00 ± 0.00	94.91 ± 0.07
im-Original	60.29 ± 13.07	97.00 ± 2.60	96.88 ± 0.07	100.00 ± 0.00	100.00 ± 0.00	85.44 ± 0.22	100.00 ± 0.00	100.00 ± 0.00	94.66 ± 0.30
im-Retrain	38.76 ± 13.41	95.86 ± 4.34	89.92 ± 5.70	24.25 ± 6.98	90.88 ± 5.03	65.22 ± 5.94	33.08 ± 11.97	95.19 ± 3.78	83.89 ± 5.83
im-Adv-R ($\mathcal{S}_r$)	39.48 ± 12.20	99.71 ± 0.19	91.04 ± 4.80	25.94 ± 6.89	96.00 ± 4.90	76.51 ± 4.32	34.76 ± 12.06	98.00 ± 4.01	87.63 ± 6.11
im-Adv-R ($\mathcal{S}_n$)	42.90 ± 11.87	97.96 ± 0.59	92.80 ± 4.54	23.97 ± 8.75	91.46 ± 1.81	67.34 ± 4.02	33.92 ± 11.85	99.37 ± 0.20	84.87 ± 5.42
im-Adv-F	64.21 ± 9.89	97.28 ± 3.34	96.81 ± 0.04	100.00 ± 0.00	100.00 ± 0.00	85.11 ± 0.21	100.00 ± 0.00	100.00 ± 0.00	94.77 ± 0.09

Adversarial unlearning helps maintain the unlearned model's utility, especially in a class-imbalanced setting

Take-Home Message

- Certification and verification are two main challenges of trustworthy machine unlearning.

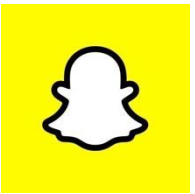
- We extended certified unlearning to non-convex models via local convex approximation and projected gradient descent.

- We proposed two adversarial unlearning methods that bypass current verification techniques, necessitating further studies on verifiable machine unlearning.

Trustworthy Machine Unlearning: Certification and Verification

Q&A

Acknowledgements



Contact: Jundong Li
https://jundongli.github.io/
jundong@virginia.edu