Trustworthy Machine Learning
Unlearning 2

Sangdon Park

POSTECH

Motivation

- Certified removal [Guo et al., 2020] assumes strongly convex loss
- Zhang et al. [2024] provides a direct extention of certified removal [Guo et al., 2020] for deep learning

Definition: Certified Unlearning

Definition ((ε, δ)-Certified Unlearning)

Let

- $\mathcal{D}$ be a training set,
- $\mathcal{D}_u \subset \mathcal{D}$ be an unlearning set,
- $\mathcal{D}_r := \mathcal{D} \setminus \mathcal{D}_u$ be a retain set,
- $\mathcal{H}$ be a hypothesis space,
- $\mathcal{A}$ be a learning algorithm.

Then, $\mathcal{U}$ is an ε-δ certified unlearning algorithm if and only if for all $\mathcal{T} \subseteq \mathcal{H}$, we have

$$\mathbb{P}\{\mathcal{U}(\mathcal{D}, \mathcal{D}_u, \mathcal{A}(\mathcal{D})) \in \mathcal{T}\} \leq e^{\varepsilon} \mathbb{P}\{\mathcal{A}(\mathcal{D}_r) \in \mathcal{T}\} + \delta$$

$$\mathbb{P}\{\mathcal{A}(\mathcal{D}_r) \in \mathcal{T}\} \leq e^{\varepsilon} \mathbb{P}\{\mathcal{U}(\mathcal{D}, \mathcal{D}_u, \mathcal{A}(\mathcal{D})) \in \mathcal{T}\} + \delta.$$

Key Theorem for Certified Unlearning

Theorem

Let

- $\tilde{w}^* := \arg\min_{w \in \mathcal{H}} \mathcal{L}(w, \mathcal{D}_r)$,
- $\tilde{w} := \mathcal{U}_{\mathtt{remove}}(w^*, \mathcal{D}_u, \mathcal{D})$, *and*
- $\|\tilde{w} - \tilde{w}^*\|_2 \leq \Delta$.

Then,

$$\mathcal{U}_{\mathtt{hide}}(w^*, \mathcal{D}_u, \mathcal{D}) := \tilde{w} + Y$$

is an ε-δ *certified unlearning if* $Y \sim \mathcal{N}(0, \sigma^2 \mathbf{I})$ *and* $\sigma \geq \frac{\Delta}{\varepsilon}\sqrt{2 \ln \frac{1.25}{\delta}}$.

- The key next step is bounding Δ.
  - With convex models, boudning Δ seems feasible.
  - How about non-convex models in general?
  - How about non-convex models in unlearning?

Algorithm: A Single-Step Newton Update ($\mathcal{U}_{\text{remove}}$)

$$\tilde{w}^* \approx \tilde{w} = w^* - H_{w^*}^{-1} \nabla \mathcal{L}(w^*, \mathcal{D}_r)$$

- $\mathcal{H}$: a set of models
- $\mathcal{D}$: an original training set
- $\mathcal{D}_u \subset \mathcal{D}$: an unlearned set
- $\mathcal{D}_r := \mathcal{D} \setminus \mathcal{D}_u$: a retained set
- $w^* := \arg\min_{w \in \mathcal{H}} \mathcal{L}(w, \mathcal{D})$: an optimal trained model – could be a local optimum
- $\tilde{w}^* := \arg\min_{w \in \mathcal{H}} \mathcal{L}(w, \mathcal{D}_r)$: an optimal unlearned model – could be a local optimum
- $\tilde{w}$: an estimated unlearned model – why? due to the Taylor expansion of $\nabla \mathcal{L}$ at w^*, *i.e.*,

$$\nabla \mathcal{L}(\tilde{w}^*, \mathcal{D}_r) \approx \nabla \mathcal{L}(w^*, \mathcal{D}_r) + H_{w^*}(\tilde{w}^* - w^*)$$

Thus, $0 = \nabla \mathcal{L}(\tilde{w}^*, \mathcal{D}_r) \approx \nabla \mathcal{L}(w^*, \mathcal{D}_r) + H_{w^*}(\tilde{w}^* - w^*)$ implies the update rule.

Main Direction

Certified Unlearning without Convexity

Bounding the approximation error $\|\tilde{w} - \tilde{w}^*\|_2$. To this end, we need following assumptions.

Assumption 1

A loss function $\ell(w, x, y)$ has an L-Lipschitz gradient in w, *i.e.*,

$$\|\nabla \mathcal{L}(w, \mathcal{D})\|_2 \leq L.$$

Assumption 2

A loss function $\ell(w, x, y)$ has an M-Lipschitz Hessian in w, *i.e.*,

$$\|H_w - H_{w'}\|_2 \leq M\|w - w'\|_2.$$

Approximation Error

Certified Unlearning without Convexity

Lemma

We have the following approximation error (given previously defined notations):

$$\|\tilde{w} - \tilde{w}^*\|_2 \leq \frac{M}{2} \|H_{w^*}^{-1}\|_2 \cdot \|w^* - \tilde{w}^*\|_2^2.$$

- Note that the proof of this lemma does not need global optimality.

Bounding the Norm of the Inverse Hessian

"Regularized" Update

$$\tilde{w} = w^* - (H_{w^*} + \lambda I)^{-1} \nabla \mathcal{L}(w^*, \mathcal{D}_r)$$

- Intuitively, we approximatly convert the non-convex objective to the strongly convex one.
  - In general, $\|H_{w^*}^{-1}\|_2$ is arbitrarly large.
  - Add a small diagonal, *i.e.*, $\|(H_{w^*} + \lambda I)^{-1}\|_2$, equivalent to the Hessian of the regularized objective, *i.e.*, $\mathcal{L}(w, \mathcal{D}_r) + \frac{\lambda}{2}\|w\|_2^2$
  - At w^*, the regularized objective can be strongly convex for some λ.
- In short, we have the following (see the paper for details):

$$\|(H_{w^*} + \lambda I)^{-1}\|_2 = \frac{1}{\lambda + \lambda_{\text{min}}}$$

  - λ_{min}: the smallest eigenvalue of H_{w^*}

Bounding the Norm of $w^* - \tilde{w}^*$

Constrained Learning

$$w^* = \arg\min_{\|w\|_2 \leq C} \mathcal{L}(w, \mathcal{D}) \quad \text{and} \quad \tilde{w}^* = \arg\min_{\|w\|_2 \leq C} \mathcal{L}(w, \mathcal{D}_r)$$

- This means that we have changed our learning algorithm to a constrained one.
- If the constraints are satisfied, we have

$$\|w^* - \tilde{w}^*\|_2 \leq \|w^*\|_2 + \|-\tilde{w}^*\|_2 \leq 2C$$

Bounding the Norm of $w^* - \tilde{w}^*$

Constrained Learning

$$w^* = \arg\min_{\|w\|_2 \leq C} \mathcal{L}(w, \mathcal{D}) \quad \text{and} \quad \tilde{w}^* = \arg\min_{\|w\|_2 \leq C} \mathcal{L}(w, \mathcal{D}_r)$$

- This means that we have changed our learning algorithm to a constrained one.
- If the constraints are satisfied, we have

$$\|w^* - \tilde{w}^*\|_2 \leq \|w^*\|_2 + \|-\tilde{w}^*\|_2 \leq 2C$$

- Can you criticize?
  - Can we actually have small C for neural networks (as $\|x\|_2$ is proportional to the dimension of x)?
  - How to find a proper C? We may implement this by regularized learning and choose to use a measured C.
  - ...

Approximation Error Bound

Main Theorem of This Paper

Theorem

With the regularized update, We have

$$\|\tilde{w} - \tilde{w}^*\|_2 \leq \left(\frac{M}{2}\|w^* - \tilde{w}^*\|_2 + \lambda\right) \|(H_{w^*} + \lambda I)^{-1}\|_2 \cdot \|w^* - \tilde{w}^*\|_2 \leq \frac{2C(MC + \lambda)}{\lambda + \lambda_{min}}.$$

- Recall that
  - $w^* = \arg\min_{\|w\|_2 \leq C}$
  - $\tilde{w} = w^* - (H_{w^*} + \lambda I)^{-1} \nabla \mathcal{L}(w^*, \mathcal{D}_r)$ with $\lambda > \|H_{w^*}\|_2$
  - $\tilde{w}^* = \arg\min_{\|w\|_2 \leq C} \mathcal{L}(w^*, \mathcal{D}_r)$
  - λ_{min}: the smallest eigenvalue of H_{w^*}
- A few notes:
  - Can we unlearn with certification from any original model?
  - Is this data-dependent bound?

Efficient Hessian Computation

Proposition

Given x i.i.d. tained samples $\{X_1, \ldots, X_s\}$, we have $\{H_{1,\lambda}, \ldots, H_{s,\lambda}\}$ of the Hessian $H_{w^*} + \lambda I$, where $H_{i,\lambda} := \nabla^2 \mathcal{L}(w^*, X_i) + \lambda I$, let

$$\tilde{H}_{i,\lambda}^{-1} = I + \left(I - \frac{H_{i,\lambda}}{H} \right) \tilde{H}_{i-1,\lambda}^{-1},$$

where $\tilde{H}_{0,\lambda}^{-1} = I$ and $\|\nabla^2 \ell(w^*, x)\| \leq H$ for all $x \in \mathcal{D}_r$. Then, $\frac{\tilde{H}_{s,\lambda}^{-1}}{H}$ is an asymptotic unbiased estimator of the inverse Hessian $(H_{w^*} + \lambda I)^{-1}$.

- Reduces sample complexity, *i.e.*, we need only s samples instead of n samples.
  - $O(np^2 + p^3) \to O(sp^2)$
- Is this effective with "data parallelization"?

Membership Inference Attack

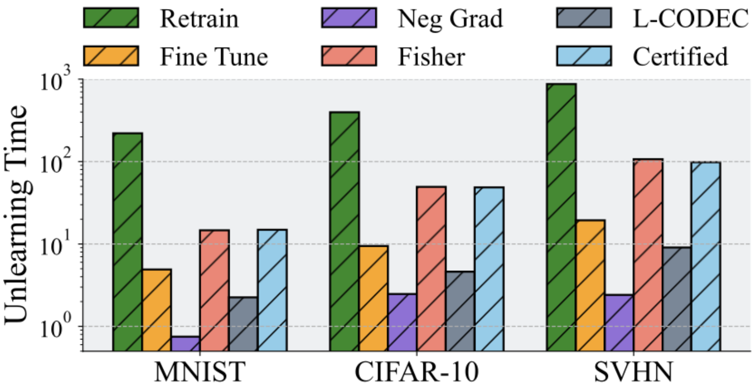
Experiment

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

- Attack Acc (= Attack F1 score) is as good as retraining.
- Here, Attack means membership inference attacks, *e.g.*,
  - For $\{(z_i, b_i)\}$ where $z_i := (x_i, y_i)$ and $b_i \in \{\text{“}z_i \notin \mathcal{D}_{\text{train}}\text{”}, \text{“}z_i \in \mathcal{D}_{\text{unlearn}}\text{”}\}$, an attacker h wins if $h(z_i)$ correctly predicts b_i

Unlearning Time

Experiment



- Efficient – note that the y-axis is log-scale.

Conclusion

- Proposes a certified unlearning method for deep models.
  - (I guess) Mainly thanks to the bounded optimal solutions, *i.e.*,

$$w^* = \arg\min_{\|w\|_2 \le C} \mathcal{L}(w, \mathcal{D}) \quad \text{and} \quad \tilde{w}^* = \arg\min_{\|w\|_2 \le C} \mathcal{L}(w, \mathcal{D}_r).$$

  - The above implies

$$\|w^* - \tilde{w}^*\|_2 \le 2C.$$

- Minimizing C is crucial for achieving a good accuracy.
  - Recall that $\min_{\|w\|_2 \le C} \mathcal{L}(w, \mathcal{D})$
  - Recall that $\sigma \ge \frac{2C(MC+\lambda)}{\varepsilon(\lambda+\lambda_{\text{min}})}\sqrt{2\ln\frac{1.25}{\delta}}$
  - Larger C $\to$ larger noise σ $\to$ accruacy drop

Reference I

C. Guo, T. Goldstein, A. Hannun, and L. Van Der Maaten. Certified data removal from machine learning models. In *International Conference on Machine Learning*, pages 3832–3842. PMLR, 2020.

B. Zhang, Y. Dong, T. Wang, and J. Li. Towards certified unlearning for deep neural networks. *arXiv preprint arXiv:2408.00920*, 2024.