Machine Unlearning Fails to Remove Data Poisoning Attacks

Anonymous Authors[1]

1. Introduction

Modern machine learning (ML) models are often trained on large-scale datasets, which include significant amounts of sensitive or personal data. This practice raises privacy concerns as the models can memorize and inadvertently reveal information about individual points in the training set. Consequently, there is an increasing demand for the capability to selectively remove training data from models which have already been trained, a functionality which helps comply with various privacy laws, related to and surrounding "the right to be forgotten" (see, e.g., the European Union's General Data Protection Regulation (GDPR) (Gen16), the California Consumer Privacy Act (CCPA), and Canada's proposed Consumer Privacy Protection Act (CPPA)). This functionality is known as *machine unlearning* (CY15), a field of research focused on "removing" specific training data points from a trained model upon request. The ideal goal is to produce a model that behaves as if the data was never included in the training process, effectively erasing all direct and indirect traces of the data. Beyond privacy reasons, there are many other applications of post-hoc model editing, including the ability to remove harmful knowledge, backdoors or other types of poisoned data, toxicity, etc.

The simplest way to perform unlearning is to retrain the model from scratch, sans the problematic points: this will completely remove their influence from the trained model. However, this is often impractical, due to the large scale of modern ML systems. Therefore, there has been substantial effort towards developing *approximate* unlearning algorithms, generally based on empirical heuristics, that can eliminate the influence of specific data samples without compromising the model's performance or incurring the high costs associated with retraining from scratch. In addition to the accuracy of the updated models, evaluation metrics, e.g. membership inference attacks (MIAs) (HSR+08; SSSS17), try to measure how much the unlearned points nonetheless affect the resulting model. Unfortunately, these evaluation metrics are themselves based on heuristics for deep learning settings, and thus success under metrics like MIAs, etc. does not guarantee that residual traces of the data do not remain, potentially allowing adversaries to recover sensitive information.

Data poisoning attacks (CGD+23; GTX+22) are a natural scenario in which the training data can have surprising and indirect effects on trained models. These attacks, now well developed for even industrial scale practical ML models, involve subtly altering a small portion of the training data, which causes the model to behave unpredictably.

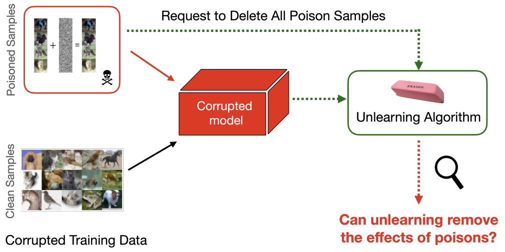


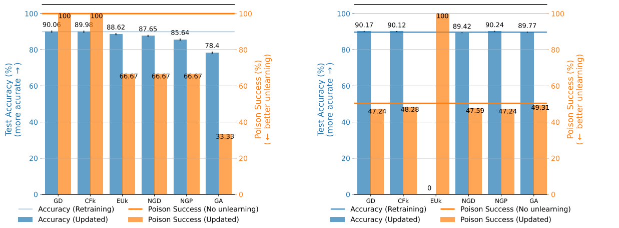
Figure 1. A corrupted ML model is trained by adding poisoned samples in the training data. In this work, we ask, whether state-of-the art machine unlearning algorithms for practical deep-learning settings can "unlearn" the effects of the poison samples.

Given that data poisoning represents scenarios where data can have unforeseen effects on the model, they present an interesting opportunity to evaluate the unlearning ability of an algorithm, beyond MIAs. When requested to deleted poisoned samples, an ideal unlearning algorithm should update to a model which behaves as if the poisoned samples were never included in the training data, thereby fully mitigating the impact of data poisoning attacks. *However, is this really the case for current unlearning methods?*

Our contribution is to show that across standard language and vision classification tasks, seven state-of-the-art unlearning algorithms consistently fail to reliably remove three different types of data poisoning from trained machine learning models. This shows that heuristic methods for unlearning can be misleading. Towards that end, our work advocated for for proper evaluations or provable guarantees as the way forward.

2. Preliminaries on Approximate Unlearning

While "retraining from scratch" on the remaining data is the gold standard for machine unlearning, this approach is generally not practically feasible for modern ML settings, e.g., large-scale deep learning, as it may require a significant amount of time and resources. Consequently, much of the research in machine unlearning has been di-



(a) Resnet-18 on CIFAR-10

(b) GPT-2 (355M) on IMDb

Figure 2. **Unlearning fails to remove targeted poisons** across a variety of unlearning methods. We poison 1.5% of the training data by adding Witch's Brew poisons (GFH+21) to a Resnet-18 trained on CIFAR-10 or instruction poisons (WWSK23) to a GPT-2 finetuned on IMDb. We then train/finetune a Resnet-18 for 100 epochs and a GPT-2 for 10 epochs on the poisoned training datasets, respectively. In both cases, we use roughly 1/10 of the original compute budget (10 epochs for CIFAR-10 or 1 epoch for IMDB) to unlearn the poisoned points. None of the considered methods remove the poisoned points.

rected towards developing approximate unlearning methods, often without rigorous theoretical guarantees, that can update a pretrained model in a computationally- and resource-efficient manner to remove the effects of a deletion set U. Some of the most popular approximate unlearning methods are: **Gradient Descent (GD)** (NRSM21), **Noisy Gradient Descent (NGD)** (CWCL24; CS23), **Gradient Ascent (GA)** (GNG21; JYY+22), **EUk** (GPS+22), **CFk** (GPS+22), **SCRUB** (KTHT24), **NegGrad+** (KTHT24).

While all of the above algorithms are designed to retain performance on the remaining training dataset and the test dataset, prior works also evaluated the unlearning capability of the above methods using various heuristics, e.g. Membership Inference Attacks, Low Memorization Accuracy, Extraction Likelihood, Interclass confusion test, etc.

3. Data Poisoning to Validate Unlearning

In a data poisoning attack, an adversary (the *attacker*) wishes to modifiy the training data provided to the machine learning model (the *victim*), in such a way that corrupted training dataset negatively impacts the model's behavior at test time. We consider the following data poisoning attacks:

- **Targeted Data Poisoning** (GFH+21): The adversary adds poison samples (by adding small perturbations to a small fraction of the training data (<1%)) to cause the model to misclassify some specific datapoints $\{(x_{\text{target}}, y_{\text{target}})\}$, from the test set S_{test}, to some pre-chosen adversarial label y_{advs}, while retaining performance on the remaining test dataset.
- **Indiscriminate Data Poisoning** (SZR+20): In an indiscriminate data poisoning attack, the adversary wishes to generate poison samples such that a model trained on the corrupted dataset S_{corr} has significantly low performance on the test dataset. The success of the attack is measured by the performance drop on the test dataset.

#Epochs	Retrain	NGP/GA	GD		CFk		EUk		SCRUB	
			1.5%	2.5%	1.5%	2.5%	1.5%	2.5%	1.5%	2.5%
2	87.04	10	83.67	83.48	68.09	59.83	29.31	25.18	83.72	82.67
4	88.23	10	85.86	85.37	69.39	61.55	39.81	33.00	85.31	83.97
6	88.79	10	86.81	86.11	70.27	62.57	43.51	38.43	85.39	84.07
8	89.14	10	87.31	86.45	70.77	63.30	47.27	40.84	85.46	84.17
10	89.24	10	87.71	86.69	71.20	63.80	49.90	43.26	85.48	84.15

Table 1. Results of unlearning indiscriminate data poisoning on CIFAR-10 in terms of test accuracy (%). The test accuracy of the poisoned models is 81.67%, 77.20%, and 69.62% for 750 and 1250 poisoned points respectively. NGP and GA exhibit random guesses (10% test accuracy) across all poison budgets. We perform a linear search for the learning rate between $[1e-6, 5e-5]$ and report the best accuracy across all methods. All the results are obtained by averaging over 8 runs.

We note that the above attacks are, so-called, *clean label* attacks, where the added perturbations are so small that a human auditor can not distinguish between poisoned and unpoisoned images. In addition to the above, we also introduce a new attack called Gaussian data poisoning attack (which generated poisons by adding i.i.d. gaussian noise and uses distribution shift to measure independence), and use it across standard language and vision classification tasks to evaluate unlearning algorithms. However, due to space constraints, we defer the detailed description and implementation details to the full version of the paper.

4. Experimental Takeaways

Key Takeaways. First none of the methods performs on par with retraining from scratch in terms of post unlearning test accuracy and effectiveness in removing the effects of data poisoning. Second, none of the evaluated methods completely removes the poisons from the models (see Figure 2, and Table 1). Third, none of the methods comes close to successfully unlearning targeted poisons from GPT-2, with some methods, such as EUk, failing entirely (see Figure 2b).

Comparing Different Unlearning Algorithms. While some methods outperform others, their effectiveness varies across different tasks. Methods like GD, CFk, and EUk typically maintain test accuracy but provide minimal to no improvement in removing targeted poisons effectively. However, in the case of indiscriminate data poisoning attacks, GD can successfully alleviate some of the poisoning effects while CFk, and EUk make the attack even stronger. Moreover, methods like NGP never come close to removing the generated poisons. Finally, GA is somewhat effective at removing targeted poisons from the Resnet-18 model, however, the test accuracy always drops by significantly more than 10% in these cases.

References

[CGD+23] Antonio Emanuele Cinà, Kathrin Grosse, Ambra Demontis, Sebastiano Vascon, Werner Zellinger, Bernhard A Moser, Alina Oprea, Battista Biggio, Marcello Pelillo, and Fabio Roli. Wild patterns reloaded: A survey of machine learning security against training data poisoning. *ACM Computing Surveys (CSUR)*, 55(13s):1–39, 2023.

[CS23] Rishav Chourasia and Neil Shah. Forget unlearning: Towards true data-deletion in machine learning. In *International Conference on Machine Learning*, pages 6028–6073. PMLR, 2023.

[CWCL24] Eli Chien, Haoyu Wang, Ziang Chen, and Pan Li. Langevin unlearning: A new perspective of noisy gradient descent for machine unlearning. *arXiv preprint arXiv:2401.10371*, 2024.

[CY15] Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *Proceedings of the 36th IEEE Symposium on Security and Privacy*, SP '15, pages 463–480. IEEE Computer Society, 2015.

[Gen16] Regulation (EU) 2016/679 of the European parliament and of the council of 27 April 2016, 2016.

[GFH+21] Jonas Geiping, Liam Fowl, W Ronny Huang, Wojciech Czaja, Gavin Taylor, Michael Moeller, and Tom Goldstein. Witches' brew: Industrial scale data poisoning via gradient matching. In *Proceedings of the 9th International Conference on Learning Representations*, ICLR '21, 2021.

[GNG21] Laura Graves, Vineel Nagisetty, and Vijay Ganesh. Amnesiac machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11516–11524, 2021.

[GPS+22] Shashwat Goel, Ameya Prabhu, Amartya Sanyal, Ser-Nam Lim, Philip Torr, and Ponnurangam Kumaraguru. Towards adversarial evaluations for inexact machine unlearning. *arXiv preprint arXiv:2201.06640*, 2022.

[GTX+22] Micah Goldblum, Dimitris Tsipras, Chulin Xie, Xinyun Chen, Avi Schwarzschild, Dawn Song, Aleksander Mądry, Bo Li, and Tom Goldstein. Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(2):1563–1580, 2022.

[HSR+08] Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V. Pearson, Dietrich A. Stephan, Stanley F. Nelson, and David W. Craig. Resolving individuals contributing trace amounts of DNA to highly complex mixtures using high-density SNP genotyping microarrays. *PLoS Genetics*, 4(8):1–9, 2008.

[JYY+22] Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. Knowledge unlearning for mitigating privacy risks in language models. *arXiv preprint arXiv:2210.01504*, 2022.

[KTHT24] Meghdad Kurmanji, Peter Triantafillou, Jamie Hayes, and Eleni Triantafillou. Towards unbounded machine unlearning. *Advances in Neural Information Processing Systems*, 36, 2024.

[NRSM21] Seth Neel, Aaron Roth, and Saeed Sharifi-Malvajerdi. Descent-to-delete: Gradient-based methods for machine unlearning. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, ALT '21. JMLR, Inc., 2021.

[SSSS17] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *Proceedings of the 38th IEEE Symposium on Security and Privacy*, SP '17, pages 3–18. IEEE Computer Society, 2017.

[SZR+20] Xu Sun, Zhiyuan Zhang, Xuancheng Ren, Ruixuan Luo, and Liangyou Li. Exploring the vulnerability of deep neural networks: A study of parameter corruption. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[WWSK23] Alexander Wan, Eric Wallace, Sheng Shen, and Dan Klein. Poisoning language models during instruction tuning. In *International Conference on Machine Learning*, pages 35413–35425. PMLR, 2023.