Article

Not peer-reviewed version

Threat Models to Machine Unlearning

Larry Milner [*]

Posted Date: 26 September 2024

doi: 10.20944/preprints202409.2068.v1

Keywords: threat model; machine unlearning; data-driven model; privacy model

Article

Threat Models to Machine Unlearning

Larry Milner

Independent Researcher, USA; quant.milner@gmail.com

Abstract: Machine unlearning is a critical process designed to allow machine learning models to forget or remove specific data upon request, particularly for privacy protection. While the primary objective of unlearning is to safeguard sensitive information, it introduces substantial privacy and security risks. This paper explores the two major categories of threats to machine unlearning: privacy attacks and security attacks. Privacy attacks, such as membership inference and model inversion, exploit residual information to infer or reconstruct sensitive data that should have been erased. Security attacks, particularly specific data poisoning, involve the injection of malicious data into the training process, which may leave lingering effects even after the unlearning process. In this paper, we provide an in-depth examination of these threats and propose several defense mechanisms. Differential privacy, adversarial training, and query limitations are highlighted as key defenses against privacy attacks, while data validation, adversarial examples, and post-unlearning audits are critical in mitigating security risks. Additionally, we discuss emerging methodologies like data provenance tracking and fine-tuning as crucial for ensuring unlearning processes are thorough and effective. Through these analyses, we aim to provide a comprehensive framework for strengthening the security and privacy of machine unlearning systems, enabling their broader adoption across industries such as healthcare, finance, and AI-driven technologies.

Keywords: threat model; machine unlearning; data-driven model; privacy model

Introduction

This section will introduce machine unlearning, its primary goal of enabling models to "forget" specific data for privacy or compliance purposes, and the rising need for privacy-preserving machine learning frameworks. It will discuss how machine learning models, once trained on data, retain and internalize knowledge that may need to be erased for privacy or regulatory compliance. This introduction will emphasize how removing or unlearning data from machine learning models is non-trivial, leading to various vulnerabilities, particularly security and privacy related. The objective of machine unlearning is to protect privacy, it might inadvertently expose the forgotten data into risk.

Literature Review

The two major threats to machine unlearning are privacy attacks and security attacks. Privacy attacks occur when adversaries exploit vulnerabilities in the unlearning process to recover or infer sensitive data that was meant to be "forgotten." One form of privacy attack is data recovery, where residual traces of deleted data remain in the model's parameters, allowing adversaries to reconstruct or reverse-engineer the forgotten information using techniques like model inversion or data reconstruction. Another common privacy threat is membership inference attacks, where an attacker can determine whether a particular data point was part of the original training set, even after the unlearning process. This exposes sensitive information about individuals or entities, undermining the very goal of machine unlearning to protect privacy. In some cases, inference leakage may occur, where unforgotten, related data can still indirectly expose information about the forgotten data.

Data recovery attack is also known as specific Data Poisoning Attack summarized by [1]. [2] utilizes the estimated covariance matrix to evaluate the impact of such attack to machine unlearning. [3] and [4] summarizes model weight gap as an effective measure before and after unlearning to assess the threat model. The major defense mechanism for such data driven attacks includes differential privacy, adversarial training and fine-tuning of the current unlearning procedures. Differential privacy was widely used in financial area as indicated by [5].

A typical privacy attack is described in the work of [6], there is a noticeable shift on the model confidence score before and after the unlearning. [7] describes another privacy attack which records model prediction label change after the unlearning. When measuring the impact of privacy attack on machine unlearning, [8] uses query comparison between unlearned and original models to gain a quantitative measure. The above three examples are membership inference attacks. The defense manner for privacy attacks is much wider. There is temperature scaling, publishing labels only, adversarial training and label smoothing. Each of these techniques helps reduce the information leakage or provide robustness to the model, ensuring that the unlearning process is more effective.

Security attacks, on the other hand, target the integrity of the unlearning process. Poisoning attacks are a major security threat, where adversaries introduce malicious data into the training process to manipulate the model's behavior. If the unlearning process does not effectively remove such poisoned data, the model remains compromised, leading to future vulnerabilities. Backdoor attacks represent another significant threat, where adversaries embed hidden triggers into models. These backdoors can remain functional even after unlearning, allowing attackers to activate malicious behavior later. Both privacy and security attacks highlight the challenges in ensuring that unlearning processes are fully effective and do not leave exploitable vulnerabilities behind.

Both privacy and security attacks underscore the complexity and challenges involved in developing effective machine unlearning algorithms. Ensuring that machine learning models can genuinely forget certain data points without leaving exploitable vulnerabilities requires ongoing research. It also necessitates the development of robust defense mechanisms, such as differential privacy, model auditing, fine-tuning unlearning procedures, and more advanced adversarial training techniques. As these threats continue to evolve, the need for sophisticated defense strategies becomes even more urgent, especially as machine unlearning finds increasing applications in privacy-critical domains such as finance, healthcare, and e-commerce.

Methodology on Privacy Attack

A privacy attack in machine unlearning occurs when an adversary tries to infer or recover sensitive information that was intended to be forgotten or erased from the model. For example, in a membership inference attack, an attacker might attempt to deduce whether a specific individual's data was used in the training set, even after that data has been supposedly unlearned. Imagine a machine learning model used by a hospital to predict disease outcomes. After a patient requests their data to be unlearned, an attacker could exploit the model's behavior to infer whether that patient's records were part of the training set by comparing its outputs before and after unlearning. This undermines the privacy guarantees, as it exposes sensitive health information that should no longer be associated with the model.

Another example is a model inversion attack, where adversaries use the model's output to reconstruct sensitive features of input data. For instance, in a facial recognition system, even after a user's data has been removed, an attacker might still be able to reconstruct their facial features by analyzing the patterns that the model retained. This is possible because some traces of the removed data might persist in the model's parameters, allowing the attacker to reverse-engineer the forgotten information.

One of the most effective defenses against privacy attacks is the use of differential privacy. This approach adds controlled noise to the data or model outputs, making it statistically impossible to distinguish whether any specific data point (such as an individual's medical record or face) was part of the training set. By obscuring individual contributions to the model, differential privacy helps ensure that membership inference attacks are thwarted.

For instance, in a financial institution's credit-scoring model, differential privacy could be employed to protect sensitive customer data, such that after a customer requests to be unlearned, no attacker could infer whether that customer's credit information was part of the original dataset.

Adversarial training is another valuable defense mechanism, where models are trained using adversarial examples—inputs designed to simulate privacy attacks. For example, a facial recognition model could be adversarial trained to resist model inversion attacks by ensuring that outputs do not reveal enough information to reconstruct sensitive features. This process teaches the model to generalize better and not leak personal information.

Other defenses include query limitation, where the number or type of queries to the model is restricted. For instance, a hospital's machine learning model could limit the number of prediction requests from any given user to prevent attackers from probing it excessively to infer information. Label smoothing and output obfuscation—where output confidence scores are deliberately softened—can also make it harder for attackers to use the model's responses to infer sensitive information, as the model's outputs become less informative.

A specific data poisoning attack occurs when an adversary intentionally introduces manipulated data into the training process to corrupt the model's behavior in a targeted way. After the unlearning process, traces of the poisoned data might still influence the model, allowing the attacker to exploit these remnants. For example, in a spam detection system, an attacker might inject data representing emails that are carefully crafted to appear harmless but are actually spam. If the unlearning process removes these spam-like emails, and the unlearning is incomplete, the attacker could still exploit weaknesses in the model to let future spam pass undetected.

Another scenario could involve financial fraud detection. Suppose an attacker introduces subtle modifications to transaction data, making fraudulent transactions resemble legitimate ones. After unlearning this data, the residual effects could persist, causing the model to miss future fraudulent activities. Worse, the attacker could manipulate the unlearning process itself by submitting requests to remove key data that was crucial to detecting fraud, thereby weakening the model's ability to protect against future attacks.

One of the primary defenses against specific data poisoning attacks is robust data validation. For instance, in a credit-scoring model, ensuring that all input data is rigorously validated for anomalies or irregularities can help prevent attackers from introducing poisoned data into the training set. This validation could include checks for inconsistent credit histories or unusually high or low transaction volumes, which might indicate an attempt to poison the model.

Additionally, adversarial training can be employed as a proactive defense. For example, in a spam detection system, adversarial examples of poisoned spam-like emails can be used during training to help the model learn how to identify and resist such attacks. By exposing the model to potential adversarial inputs, the system becomes more resilient to future poisoning attempts.

After the unlearning process, post-unlearning audits should be conducted to ensure that no traces of the poisoned data remain. This involves evaluating the model's behavior and performance after unlearning, to ensure that it does not exhibit any biases or vulnerabilities that could have been introduced by the poisoned data. For instance, in a financial fraud detection system, such audits would monitor the model's ability to detect fraud after the removal of suspicious data, ensuring that the unlearning process was effective.

Another important defense is data provenance tracking, which involves maintaining a record of the origins and transformations of all data points used in training. This helps identify and isolate any potentially poisoned data before it can influence the model. In a healthcare model predicting patient outcomes, data provenance tracking would ensure that data from unreliable or suspicious sources is flagged before it can corrupt the model.

Finally, employing regularization techniques such as fine-tuning the model after unlearning can mitigate the effects of any residual poisoned data. For instance, in a system predicting customer churn, fine-tuning the model after unlearning corrupted customer data would help to adjust the model's parameters, reducing the chance that poisoned information could influence future predictions.

Methodology on Security Attacks

Security attacks for machine unlearning includes malicious unlearning request attack, data reconstruction attack, and jailbreak attack. We try our best to find literature that introduce security attacks on machine unlearning.

1. Malicious Unlearning Request Attack

A malicious unlearning request attack involves an adversary exploiting the unlearning process itself by submitting false or deceptive requests to force the model to forget important or benign data, thus causing it to underperform or behave abnormally. These types of attacks target the unlearning mechanism that is designed to remove certain data points from a model. However, if an attacker can manipulate this mechanism, they can misuse the unlearning process to achieve malicious ends, such as weakening the model's ability to make accurate predictions or disrupting its operational effectiveness. [9] is an attack with poisoned training dataset that involves unexpected unlearning request attack. The original model cannot differentiate the valid request and the malicious one.

In malicious unlearning request attacks, adversaries typically pose as legitimate users who wish to remove data from the model, but their true goal is to remove strategically important data or otherwise undermine the model's functionality. This can be accomplished in several ways: 1. Overloading the unlearning system with an excessive number of unlearning requests can cause the model to experience a performance degradation, as repeated requests for unlearning certain data points could impact the model's ability to generalize effectively. 2. Selective unlearning of critical data: An attacker might target specific, highly informative data points for unlearning. By removing or forgetting highly influential data, the model may become less accurate, biased, or incapable of responding appropriately to certain inputs.

An example of a malicious unlearning request attack could occur in the context of a recommender system that uses machine learning models to predict user preferences. An adversary could submit numerous unlearning requests to delete the data related to popular items or preferences, thereby skewing the recommendations toward lesser-known or even harmful items. In another instance, a malicious actor might target the unlearning of critical data in a healthcare model that predicts patient diagnoses, causing the model to lose valuable diagnostic insights.

In terms of defense against such an attack, for instance, imagine a fraud detection system designed to detect fraudulent financial transactions. An attacker who successfully requests unlearning of key fraudulent transaction patterns could cause the system to forget these crucial patterns, leaving it unable to detect future fraud. This not only weakens the model's ability to catch fraudulent activities but also opens the door for the adversary to launch further financial attacks undetected.

To defend against this type of attack, it is essential to verify the authenticity of unlearning requests by employing robust identity verification processes and auditing systems that track the legitimacy of the unlearning request. Techniques such as request authentication, limiting the number of unlearning requests that a user can submit in a certain time period, and monitoring patterns of suspicious unlearning requests can help mitigate this risk. Moreover, incorporating data importance weighting mechanisms can ensure that highly critical or valuable data is protected from unlearning without rigorous verification [10].

2. Data Reconstruction Attack

A data reconstruction attack is a highly sophisticated security threat in which adversaries attempt to recover forgotten data after it has been removed from the model. Despite efforts to unlearn specific data, the model may still retain subtle traces or patterns that allow an attacker to reconstruct the original data with sufficient accuracy. This type of attack capitalizes on the fact that many machine learning models, especially complex ones like deep neural networks, may not completely "forget" certain information even after the unlearning process has been applied.

Data reconstruction attacks typically rely on model inversion or gradient-based methods to uncover information that has been theoretically unlearned. By probing the model with specific inputs or by analyzing the gradients during the learning process, attackers can infer how certain data influenced the model's parameters, even if the data was supposedly forgotten. This allows adversaries to reverse-engineer the underlying patterns, relationships, or even the exact data points that were used to train the model. [11] describes a carefully designed unlearned data request that passes the security defense. [12] introduces a similar crafted request that passes both the security and membership verification. They suggest that hash verification and membership verification are both needed to counter such attacks.

For example, in a machine learning model trained on facial recognition data, an attacker might exploit a data reconstruction attack to infer facial features or even generate an image that closely resembles a person whose data was meant to be unlearned. Similarly, in financial models, an attacker could reconstruct key financial records or transaction details that were theoretically removed, leading to severe privacy breaches and security vulnerabilities.

An illustrative example of a data reconstruction attack can be found in medical data systems. Suppose a model is trained on patient health data to predict disease outcomes. Even after a patient request that their data be removed through machine unlearning, an adversary could employ data reconstruction techniques to recover sensitive health information. This could include inferring the patient's disease history or even identifying specific individuals, despite their data being "forgotten."

Defending against data reconstruction attacks requires careful design of the unlearning process. Differential privacy is a critical defense mechanism that can be employed to ensure that the removed data cannot be reconstructed from the remaining information. Additionally, model auditing tools that analyze the extent of forgetting after unlearning can be used to assess whether the unlearning process was fully effective. Implementing model regularization techniques can also reduce the amount of residual information left behind after unlearning, making it more difficult for adversaries to reconstruct forgotten data.

3. Jailbreak Attack

A jailbreak attack targets the internal constraints or limitations imposed on machine learning models, often aimed at exploiting the model's ability to perform unauthorized or unintended actions. In the context of machine unlearning, a jailbreak attack involves bypassing the protections designed to ensure that the model has forgotten specific data. This attack can be viewed as an effort to "unlock" or regain access to information that has been removed or restricted by the unlearning process. In such cases, attackers may find ways to extract forgotten data or regain control over model behaviors that were disabled through unlearning. [13] describes such an attack during machine unlearning process that is embedded in in-context learning.

In a jailbreak attack, the adversary seeks to manipulate the model's operational constraints by exploiting weaknesses in the unlearning process. This could be achieved through various means: 1. Bypassing access controls or security mechanisms designed to enforce unlearning. An attacker could attempt to trick the system into ignoring unlearning constraints, effectively restoring access to forgotten data. 2. Activating hidden functionalities or bypassing input validation checks that were disabled during unlearning. By sending crafted inputs, an adversary can cause the model to behave in ways it was not intended to after the unlearning process.

For instance, in a security system based on machine learning that has undergone unlearning to remove certain privileges or access patterns, a jailbreak attack could enable the adversary to bypass these restrictions and regain unauthorized access to protected areas of the system.

[14] and [15] illustrate jailbreak attacks to machine unlearning, both suggest the increase of adversarial mechanisms as the ultimate defense solution. An example of a jailbreak attack could occur in the context of AI systems used in autonomous vehicles. Suppose a model has been unlearned to disable certain driving behaviors deemed unsafe, such as speeding in restricted areas. A skilled adversary might launch a jailbreak attack to re-enable these behaviors, allowing the

vehicle to disregard speed limits or other safety constraints. In another example, an attacker might bypass the unlearning protections in a machine learning system that controls access to sensitive databases, enabling them to recover previously restricted or forgotten data.

To defend against jailbreak attacks, it is important to implement strong security policies that prevent the reactivation of unlearned behaviors or the restoration of forgotten data. Input validation and strict access control mechanisms can help ensure that unauthorized actions cannot be triggered. Moreover, continuous monitoring of model outputs and auditing of system behaviors after unlearning is necessary to detect and mitigate potential jailbreak attempts. By incorporating robust logging and anomaly detection systems, administrators can quickly identify when an adversary is attempting to circumvent the constraints imposed by unlearning.

Conclusion

Machine unlearning holds significant promise in the field of privacy protection by allowing models to forget data that individuals or entities request to be removed. However, this process introduces both privacy and security risks that must be carefully managed. Privacy attacks, such as membership inference and model inversion, reveal how residual data traces can be exploited by adversaries to infer sensitive information. Specific data poisoning attacks highlight how attackers can deliberately corrupt models by introducing malicious data that may continue to affect the model's behavior even after the unlearning process.

Addressing these risks requires implementing robust defense mechanisms tailored to the nature of the threat. Techniques such as differential privacy, adversarial training, and query limitations are essential to protecting against privacy attacks. For specific data poisoning attacks, strategies like data validation, adversarial examples, and post-unlearning audits are critical for ensuring that unlearning effectively removes both data and its harmful influence. In particular, integrating data provenance tracking and fine-tuning models post-unlearning can help safeguard systems from residual effects left by malicious actors.

As machine unlearning becomes more prevalent in fields such as healthcare, finance, and AI-driven industries, the challenges outlined in this paper emphasize the need for continuous research and innovation. Developing stronger defenses and refining unlearning methodologies are crucial steps toward mitigating the inherent vulnerabilities of machine unlearning and ensuring that models remain secure, reliable, and privacy conscious. By addressing both privacy and security attacks, organizations can confidently deploy machine unlearning in a way that meets the growing demands for data privacy while preserving the integrity and functionality of their machine learning systems.

References

1. N. Li, C. Zhou, Y. Gao, H. Chen, A. Fu, Z. Zhang and Y. Shui, "Machine Unlearning: Taxonomy, Metrics, Applications, Challenges, and Prospects," *arXiv preprint arXiv:2403.08254*, 2024.
2. M. Bertran, S. Tang, M. Kearns, J. Morgenstern, A. Roth and Z. S. Wu, "Reconstruction Attacks on Machine Unlearning: Simple Models are Vulnerable," *arXiv preprint arXiv:2405.20272*, 2024.
3. J. Du, Z. Wang and K. Ren, "Textual Unlearning Gives a False Sense of Unlearning," *arXiv preprint arXiv:2406.13348*, 2024.
4. Z. Wang, Y. Zhu, Z. Li, Z. Wang, H. Qin and X. Liu, "Graph neural network recommendation system for football formation," *Applied Science and Biotechnology Journal for Advanced Research*, vol. 3, p. 33–39, 2024.
5. Z. Li, B. Wang and Y. Chen, "Incorporating economic indicators and market sentiment effect into US Treasury bond yield prediction with machine learning," *Journal of Infrastructure, Policy and Development*, vol. 8, p. 7671, 2024.
6. M. Chen, Z. Zhang, T. Wang, M. Backes, M. Humbert and Y. Zhang, "When machine unlearning jeopardizes privacy," in *Proceedings of the 2021 ACM SIGSAC conference on computer and communications security*, 2021.
7. Z. Lu, Y. Wang, Q. Lv, M. Zhao and T. Liang, "FP 2-MIA: A Membership Inference Attack Free of Posterior Probability in Machine Unlearning," in *International Conference on Provable Security*, 2022.
8. H. Hu, S. Wang, T. Dong and M. Xue, "Learn what you want to unlearn: Unlearning inversion attacks against machine unlearning," *arXiv preprint arXiv:2404.03233*, 2024.

doi:10.20944/preprints202409.2068.v1

9. J. Z. Di, J. Douglas, J. Acharya, G. Kamath and A. Sekhari, "Hidden poison: Machine unlearning enables camouflaged poisoning attacks," in *NeurIPS ML Safety Workshop*, 2022.
10. Y. Wei, X. Gu, Z. Feng, Z. Li and M. Sun, "Feature Extraction and Model Optimization of Deep Learning in Stock Market Prediction," *Journal of Computer Technology and Software*, vol. 3, 2024.
11. H. Hu, S. Wang, J. Chang, H. Zhong, R. Sun, S. Hao, H. Zhu and M. Xue, "A duty to forget, a right to be assured? exposing vulnerabilities in machine unlearning services," *arXiv preprint arXiv:2309.08230*, 2023.
12. C. Zhao, W. Qian, R. Ying and M. Huai, "Static and sequential malicious attacks in the context of selective forgetting," *Advances in Neural Information Processing Systems*, vol. 36, 2024.
13. I. Shumailov, J. Hayes, E. Triantafillou, G. Ortiz-Jimenez, N. Papernot, M. Jagielski, I. Yona, H. Howard and E. Bagdasaryan, "UnUnlearning: Unlearning is not sufficient for content regulation in advanced generative AI," *arXiv preprint arXiv:2407.00106*, 2024.
14. H. Yuan, Z. Jin, P. Cao, Y. Chen, K. Liu and J. Zhao, "Towards Robust Knowledge Unlearning: An Adversarial Framework for Assessing and Improving Unlearning Robustness in Large Language Models," *arXiv preprint arXiv:2408.10682*, 2024.
15. L. Schwinn, D. Dobre, S. Xhonneux, G. Gidel and S. Gunnemann, "Soft prompt threats: Attacking safety alignment and unlearning in open-source llms through the embedding space," *arXiv preprint arXiv:2402.09063*, 2024.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.