Critique of Unlearning Inversion Attacks against Machine Unlearning

Farah Zaghdane

June 2024

1 Introduction

The "*Learn What You Want to Unlearn:Unlearning Inversion Attacks against Machine Unlearning*" paper [1] by *Hongsheng Hu*, *Shuo Wang*, *Tian Dong* and *Minhui Xue*, presents a novel investigation into the extent to which machine unlearning methods can leak information about the unlearned data. It proposes two types of inversion attacks, **feature recovery** and **label inference**, that can exploit the privacy vulnerabilities in machine unlearning techniques. The authors demonstrate that by having access to the original and unlearned models, an adversary can recover sensitive information about the unlearned data, such as its features and labels. This is a significant finding as machine unlearning is intended to protect the privacy of individuals whose data is removed from the model.

2 Problem and Literature Review

The problem statement clearly articulates the research objective, which is to study the information leakage in machine unlearning methods. The authors provide relevant background information and a comprehensive review of the literature, demonstrating a strong understanding of the existing research in this area. The review critically analyzes the current state of machine unlearning and highlights the need for a deeper investigation into the privacy implications of these techniques, forming a solid rationale for the current study.

3 Hypotheses and Methodology

The paper presents specific hypotheses to be tested, focusing on the ability of unlearning inversion attacks to recover the feature and label information of unlearned data. The hypotheses are clearly stated and testable.

The methodology section provides a detailed description of the experiments, including the subjects, instruments, and procedures. The authors used deep neural networks (DNNs) trained on the CIFAR-10, CIFAR-100, Chest X-Ray, German Credit, and ImageNet datasets, which are common benchmarks in the field. The sample size and selection method are described, though the generalizability of the findings to other model architectures and datasets remains a limitation.

The instruments, including the machine unlearning methods (**exact unlearning** via retraining and **approximate unlearning** via single gradient update) and the attack algorithms, are described in detail.

Overall, the methodology section presents a thorough description of the experimental setup and evaluation, though the lack of a comprehensive assessment of the instrument validity is noted as a potential limitation.

4 Results and Discussion

The results section presents empirical findings on the efficacy of the proposed unlearning inversion attacks in recovering feature and label information from unlearned models. Appropriate statistical analyses were employed to quantify the performance of the attacks, with the feature recovery attack achieving low MSE and high PSNR, and the label inference attack attaining notable classification accuracy.

In the discussion, the authors contextualize these findings within the existing literature, discuss the theoretical and practical implications, and identify three potential defenses - **parameter obfuscation**, **model pruning**, and **fine-tuning** - evaluating their effectiveness in mitigating the attacks.

They interpret the results as evidence of significant privacy vulnerabilities in machine unlearning, as adversaries can exploit access to the original and unlearned models to recover sensitive information.

5 Limitations and Future Directions

While the paper makes a valuable contribution to the understanding of privacy risks in machine unlearning, the study has some limitations that should be addressed in future research:

5.1 Scope of Machine Unlearning Inversion Attacks Considered

The primary focus of the paper was on unlearning inversion attacks that target the recovery of feature and label information from the unlearned data. While this is an important aspect of privacy preservation, it represents only a subset of the potential attack vectors that can compromise the security of machine unlearning.

The authors did not explore attacks on the model parameters or architecture, which could potentially reveal additional sensitive information about the unlearned data. For instance, adversaries could misuse the model parameters to retain traces of the original training data, even after unlearning. Alternatively, they could utilize the model architecture as a side-channel to infer details about the original training data.

Furthermore, the paper did not consider the possibility of compound attacks, where multiple attack strategies are combined to achieve a more comprehensive breach of privacy.

To fully understand the security landscape of machine unlearning, future research should expand the scope of investigated attacks to cover a wider range of attack vectors and their potential interactions.

5.2 Generalizability of Attacks Across Machine Learning Models, Datasets, and Application Domains

The paper focuses on machine learning classification tasks, specifically on deep neural network (DNN) models. While this is a common and important application, the proposed attacks may not generalize well to other machine learning tasks, such as regression or unsupervised learning.

The experiments are conducted on a limited set of benchmark datasets (CIFAR-10, CIFAR-100, STL-10) and model architectures (ConvNet, ResNet-18). It is unclear how the attacks would perform on a more diverse range of datasets and model types, which could have different properties and vulnerabilities.

To enhance the practical relevance of the research, future studies should evaluate the unlearning inversion attacks on a broader range of model architectures, datasets, and application domains. This will provide a more comprehensive understanding of the attack's generalizability and inform the development of robust unlearning techniques that can withstand a wider spectrum of privacy threats.

5.3 Practical Feasibility of Attacks

The paper's assumption that the attacker has either white-box access to both the original and unlearned models (for the feature recovery attack) or black-box access (for the label inference attack) simplifies the attack scenario. However, in reality, obtaining such unrestricted access to the target models may be challenging, as machine learning models are often deployed as part of larger, complex systems with various security measures in place.

Attackers may encounter difficulties in acquiring the necessary credentials, computational resources, or network access required to execute the proposed unlearning inversion attacks.

Furthermore, the paper did not consider the time and effort required by the attacker to set up and execute the attacks. In a practical setting, the feasibility of the attacks may be heavily influenced by the attacker's resources, technical expertise, and the overall cost-benefit analysis of the attack.

Future research should investigate the practical limitations and challenges faced by attackers in obtaining the required access and resources. This will provide a more realistic assessment of the actual threats posed by

such attacks and guide the development of appropriate countermeasures that address real-world constraints faced by potential adversaries. The analysis should incorporate more realistic access constraints for the attacker, such as partial or restricted access to the target models.

5.4 Machine Unlearning Methods Considered

The paper under critique focuses on evaluating the proposed unlearning inversion attacks against two specific machine unlearning methods: exact unlearning (also known as retraining) and approximate unlearning (using a single gradient update).

Exact unlearning refers to the process of completely retraining the machine learning model from scratch, excluding the data that needs to be removed. This approach guarantees the complete removal of the target data from the model, but can be computationally expensive, especially for large-scale models and datasets.

In contrast, approximate unlearning aims to achieve a similar effect to exact unlearning, but with reduced computational cost. The single gradient update approach evaluated in the paper modifies the model parameters by applying a single gradient step in the direction opposite to the gradient computed on the data to be unlearned. This technique is more efficient than retraining the entire model, but may not provide the same level of guarantee for complete data removal.

While the paper's focus on these two unlearning methods is understandable, given their widespread use and practical relevance, it is important to note that there are other machine unlearning techniques [2] that were not considered in this study. These include, but are not limited to, influence function-based methods [3], Shapley value-based approaches [4], and more advanced techniques that leverage additional information, such as the training data distribution or the model's internal structure.

Expanding the evaluation to include a broader range of machine unlearning methods would provide a more comprehensive understanding of the privacy vulnerabilities associated with machine unlearning. Assessing the performance of the proposed unlearning inversion attacks against a diverse set of machine unlearning techniques would help identify the fundamental limitations of current unlearning approaches and guide the development of more robust and secure machine unlearning solutions.

6 Conclusion

The research on unlearning inversion attacks presented in the original paper has made valuable contributions to the understanding of privacy risks in machine unlearning. However, the limitations identified in this critique, such as the narrow scope of attacks considered, the lack of generalizability, and the practical feasibility concerns, highlight the need for further research to develop a more comprehensive and realistic understanding of the security implications of machine unlearning.

By addressing these limitations and expanding the scope of investigation, future studies can provide a more robust and holistic assessment of the privacy vulnerabilities in machine unlearning and inspire the development of more secure and resilient unlearning techniques. This will be crucial as machine learning systems become increasingly pervasive, and the need for effective and privacy-preserving unlearning mechanisms continues to grow.

References

[1] Hongsheng Hu et al. *Learn What You Want to Unlearn: Unlearning Inversion Attacks against Machine Unlearning*. 2024. arXiv: 2404.03233 [cs.CR]. URL: https://arxiv.org/abs/2404.03233.

[2] Jie Xu et al. "Machine Unlearning: Solutions and Challenges". In: *IEEE Transactions on Emerging Topics in Computational Intelligence* 8.3 (June 2024), pp. 2150–2168. ISSN: 2471-285X. DOI: 10.1109/tetci.2024.3379240. URL: https://arxiv.org/abs/2308.07061.

[3] Pang Wei Koh and Percy Liang. *Understanding Black-box Predictions via Influence Functions*. 2020. arXiv: 1703.04730 [stat.ML]. URL: https://arxiv.org/abs/1703.04730.

[4] Ruoxi Jia et al. *Towards Efficient Data Valuation Based on the Shapley Value*. 2023. arXiv: 1902.10275 [cs.LG]. URL: https://arxiv.org/abs/1902.10275.