机器遗忘学习的自身安全问题研究

谭尚谋

2118021-21009201012

摘 要 机器遗忘学习（Machine Unlearning）是近年来为满足用户数据隐私需求而提出的重要方法。通过主动删除模型中的特定信息，实现了对敏感数据的快速移除。然而，这一领域存在显著的安全问题，例如信息泄露、攻击面增大以及防御机制的不足。本文综述了机器遗忘学习的典型漏洞，包括模型重建攻击、对抗样本生成和遗忘验证的潜在问题，并探讨了相关的防御措施及其局限性。最后，本文提出了一些未来的研究方向，以期提升机器遗忘学习的安全性和实用性。

关键词 机器遗忘学习，信息安全，漏洞，防御机制，隐私保护

Machine Unlearning Security Issues Research

Tan Shangmou

Abstract Machine unlearning (MU) has emerged as an important method to meet user data privacy needs in recent years. By actively deleting specific information in the model, it achieves rapid removal of sensitive data. However, this field has significant security issues, such as information leakage, increased attack surfaces, and insufficient defense mechanisms. This paper reviews the typical vulnerabilities of machine unlearning, including model reconstruction attacks, adversarial sample generation, and potential problems with forgetting verification, and discusses related defense measures and their limitations. Finally, this paper proposes some future research directions to improve the security and practicality of machine unlearning.

Key words Machine Unlearning, Information Security, Vulnerabilities, Defense Mechanisms, PrivacyProtection

1 引言

机器遗忘学习（Machine Unlearning）作为一项新兴的研究领域，旨在满足用户对数据隐私保护的需求，其核心目标是通过在训练后有效删除特定数据的影响，保证这些数据对模型的贡献被彻底移除。这一技术的重要性在于响应用户的隐私权需求，特别是在数据法规日益严格的背景下，例如《通用数据保护条例》（GDPR）[1]中“被遗忘权”的规定。然而，尽管机器遗忘学习具有显著的实用价值，其实现过程仍然面临着多方面的技术挑战和潜在的安全隐患。在遗忘过程中，模型的整体性能可能会显著下降，尤其是在遗忘的数据对模型决策起到关键作用时，这种性能下降更为明显[2]。此外，确保指定数据确实被彻底遗忘也是一大挑战，设计有效的验证方法以判断数据是否完全移除在理论和实践中都存在难度。同时，遗忘过程可能会引发攻击者的逆向推断风险，例如通过模型输出变化或梯度信息推测被遗忘的数据特性，进而破坏隐私保护目标。计算代价也是不可忽视的问题，特别是对深度

提交作业日期：年-月-日；
*

学习模型而言，重新训练以移除特定数据会带来巨大的资源消耗，这在大规模数据和模型场景中尤为突出。此外，由于算法局限性或实现错误，遗忘可能是不完全的，某些特征仍可能隐含在模型参数中，为潜在攻击提供了可能性。针对这些问题，研究者提出了一系列改进方向，例如通过优化的遗忘算法减少对模型性能的影响[3]，采用数学证明或实验验证提高遗忘效果的验证方法，结合噪声注入或差分隐私技术增强对逆向攻击的防御，设计智能化的数据选择策略减少潜在隐私风险[4]，在联邦学习框架下实现分布式遗忘以降低集中处理带来的问题，以及推动开源工具和标准化实践以提高研究透明性和实用性。综上所述，机器遗忘学习作为保护用户隐私的重要技术[5]，既为解决数据隐私问题提供了可能性，也带来了新的挑战[6][7]。在未来的研究中，通过持续优化算法、验证机制和安全策略，这一领域有望在实际应用中得到更广泛的部署，但同时需要关注隐私保护与模型性能之间的平衡，以确保其在复杂场景下的可靠性和实用性。

2. 机器遗忘学习的基本原理

机器遗忘学习是一种新兴的技术，旨在通过技术手段从训练好的模型中移除特定数据的影响，以保护用户隐私。这一过程不仅响应了对数据隐私保护的需求，尤其是在《通用数据保护条例》（GDPR）等法规背景下，更是对“被遗忘权”的实际技术实现。[8]

2.1 定义与背景

机器遗忘学习的核心在于，它允许我们在模型训练完成后，有效地删除特定数据的影响，确保这些数据对模型的贡献被彻底移除。这与传统的数据删除不同，后者需要从源头删除数据，而机器遗忘学习则关注于模型层面的数据影响消除。

2.2 基于梯度更新的遗忘方法

基于梯度更新的遗忘方法是机器遗忘学习中的一种技术手段[9]。这种方法的原理在于直接调整模型参数，使其偏离被遗忘数据对模型的影响。在深度学习模型中，参数是通过训练数据优化的结果，而基于梯度的方法则利用这一机制，通过计算遗忘数据对模型参数的贡献（如梯度方向和大小），然后在参数空间内“反向抵消”其影响。这种方法的实现细节包括参数逆向调整、递归梯度更新以及显式正则化，以平衡遗忘效率与模型性能[10]。

应用场景包括在线学习模型，这些模型需要动态更新数据的影响，例如推荐系统中的实时数据更新。此外，当部分数据被标记为误差时，基于梯度更新的遗忘方法可以删除错误数据对模型的影响，提升模型精度。这种方法的优点在于遗忘过程快速，适合实时场景，并且与现有的优化算法（如SGD）兼容，易于实现。然而，对于深度模型而言，计算残余梯度的复杂性较高，可能遗留部分信息，且在数据量较大时，逆向梯度计算成本显著增加。

2.2.1 基于模型重构的遗忘方法

基于模型重构的遗忘方法则采取了一种更为彻底的遗忘策略[11]，即通过重新构建模型的结构或优化过程，从根本上移除敏感数据的影响。这种方法的实现细节包括子模型提取、知识蒸馏和模型剪枝。子模型提取涉及分析模型参数与数据特征的关系，从模型中提取与敏感数据无关的部分，形成“子模型”。知识蒸馏则是将原模型在非敏感数据上的行为提取为“软标签”（soft labels），并训练一个新模型来模仿这些行为，目标是尽可能保留原模型对非敏感数据的预测能力。模型剪枝通过剪枝操作移除包含敏感数据信息的部分网络结构或节点。

基于模型重构的遗忘方法的应用场景包括数据隐私要求极高的系统，如医疗诊断模型，以及深度学习中需要高精度遗忘的场景，如金融模型中的敏感交易数据[12]。这种方法的优点在于遗忘效果彻底，敏感数据从理论上不可恢复，适用于多种模型类型，包括深度神经网络和传统机器学习模型。然而，重构过程可能导致显著的计算开销，尤其在模型规模较大时，蒸馏过程可能导致部分性能损失，特别是原模型过于复杂的情况下。

3. 机器遗忘学习的安全问题

机器遗忘学习在保护隐私方面具有重要作用[13]，但其实施过程中也可能引发新的安全问题，这些问题不仅威胁模型安全，还可能导致隐私泄露和性能下降。本文详细探讨了信息泄露风险、对抗性攻击面增加以及遗忘验证的技术挑战，并结合实际示例进行说明。

3.1 信息泄露风险

遗忘操作可能导致模型的敏感参数被攻击者利用，从而泄露隐私信息。这种风险主要源于遗忘过程未能彻底移除敏感数据的影响，或在移除过程中暴露了模型内部的细节。例如，模型参数可能仍然包含被遗忘数据的部分特征，或者攻击者通过分析模型的输出或决策边界，逆向推断出遗忘数据的特性[14]。为了防御这类风险，可以采用差分隐私技术在训练和遗忘过程中引入噪声[15]，降低敏感数据的直接映射关系，或者对敏感数据影响较大的参数进行随机化或初始化，以及限制攻击者对模型参数或输出的访问权限。

3.2 对抗性攻击面

遗忘后的模型可能因参数调整或结构变化而引入新的弱点，使其更易受到对抗性攻击。攻击者可以生成伪造数据来误导模型，导致预测错误。例如，利用模型的梯度信息生成小幅扰动样本，使模型的预测结果发生显著变化，或者训练一个伪造模型并利用其对抗样本攻击遗忘后的目标模型[16]。为了减轻这种威胁，可以通过对抗训练在模型训练和遗忘过程中加入对抗样本，使模型对对抗性扰动更具鲁棒性，或者通过检测机制过滤潜在攻击样本，以及定期评估遗忘后模型的对抗性弱点，并采取修复措施。

遗忘验证问题也是机器遗忘学习中的一个核心挑战。由于模型的高复杂性和参数的非线性关系，传统验证方法难以量化遗忘的完整性。遗忘验证的技术挑战包括如何衡量遗忘数据在模型参数或输出中的残留程度，以及如何提高验证过程的效率。为了解决这些问题，可以采用基于重构的验证方法，通过重构模型的子结构验证遗忘数据的影响是否被移除，或者利用差分隐私或信息熵计算残留信息的概率分布，以及引入独立的验证机构或工具，确保验证过程的公正性和准确性。

4 防御措施与研究进展

为应对机器遗忘学习中的安全问题，研究人员提出了多种防御措施，并对遗忘效率、验证可信度以及对抗性攻击防御进行了深入探索。本节将从三个主要方面详细阐述相关进展。

4.1 提升遗忘效率的算法设计

在机器遗忘学习中，提升遗忘操作的效率至关重要，尤其是在面对大规模数据或深度模型时。增量式训练优化成为当前研究的重点方向。

通过将遗忘过程与增量式训练相结合，只更新受影响的参数，而非重新训练整个模型。

通过数据分片训练（Data Sharding Training）将数据集划分为多个独立分片，并针对需要遗忘的数据分片执行增量更新。例如，“Selective Influence Subset Aggregation”（SISA）方法通过分片训练实现快速遗忘。在训练过程中缓存梯度更新信息，仅对受遗忘数据影响的参数进行调整。针对深度模型，仅对与遗忘数据强相关的网络层参数进行微调，而保留其他层的参数。在推荐系统中，当用户请求删除其行为数据时，通过增量训练，仅更新相关推荐模型的参数，而不重新训练整个系统。在在线广告投放中，动态移除特定用户行为数据，实现隐私保护的同时保持系统实时性。

4.2 改进遗忘验证的可信度

验证遗忘是否彻底且可信是机器遗忘学习中的核心问题。近年来，引入可验证计算（Verifiable Computation）为遗忘验证提供了新方向[17]。如可验证计算的引入，通过密码学或形式化验证技术，确保遗忘操作的正确性，并提供可验证的证据。

4.3 减少对抗性攻击影响

对抗性攻击对经过遗忘处理的模型构成重大威胁。为了减轻这种威胁，鲁棒性优化方法显得尤为重要[18]，它们通过优化模型的训练和遗忘流程，增强模型抵御对抗性攻击的能力。在遗忘数据的同时，引入对抗样本进行协同训练，能够提升模型对于对抗性扰动的鲁棒性[19]。例如，可以采用投影梯度下降（PGD）算法[20]来生成对抗样本，并将其纳入训练过程中。在提升遗忘效率方面，我们可以通过增量式训练和精细化的优化策略，实现快速且高效的遗忘操作，从而减少对模型性能的影响；在提高遗忘验证的可信度方面，引入了密码学和形式化验证技术，确保遗忘操作的完整性和透明度，为遗忘操作的正确性提供坚实的证据支持。为了降低对抗性攻击的影响，我们利用鲁棒性优化技术，增强遗忘后模型的安全性和防御能力，确保模型在面对潜在攻击时仍能保持稳定性和可靠性。

5 总结

本文深入探讨了机器遗忘学习（Machine Unlearning）的安全性问题，这是一项为满足用户数据隐私需求而发展的重要技术。文章首先概述了机器遗忘学习的基本原理，包括其定义、背景以及基于梯度更新和模型重构的遗忘方法。随后，文章详细分析了机器遗忘学习过程中可能遇到的安全问题，如信息泄露风险、对抗性攻击面的增加，以及遗忘验证的技术挑战，并提出了相应的防御措施，包括提升遗忘效率的算法设计、改进遗忘验证的可信度，以及减少对抗性攻击影响的策略。

尽管机器遗忘学习在保护用户隐私方面具有重要意义，但其实施过程中也可能带来新的安全威胁。为了应对这些挑战，研究人员提出了多种防御措施，并探讨了其局限性。文章旨在提升机器遗忘学习的安全性和实用性，确保在保护隐私的同时，也能维护模型的性能和可靠性。通过这些研究，机器遗忘学习有望在实际应用中得到更广泛的部署，为数据隐私保护提供有力支持。

参 考 文 献

[1] Mojtaba Valipour, Bowen You, Maysum Panju, Ali Ghodsi 。SymbolicGPT: A Generative Transformer Model for Symbolic Regression https://arxiv.org/abs/2106.14131

[2] Veldanda, S. P., & Domingos, P. (2021). Machine Unlearning: A Survey. *arXiv preprint arXiv:2103.03279*.

[3] Ginart, A., Guan, M. Y., Valiant, G., & Zou, J. Y. (2019). Making AI Forget You: Data Deletion in Machine Learning. *Advances in Neural Information Processing Systems*, 32.

[4] Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., et al. (2021). Machine Unlearning. *IEEE Symposium on Security and Privacy (SP)*, 2021.

[5] Liu, Y., Shen, H., & Zhang, J. (2023). Adversarial Robustness of Machine Unlearning. *Proceedings of the AAAI Conference on Artificial Intelligence*.

[6] Tom B. Brown, Dandelion Mané, Aurko Roy, Martín Abadi, Justin Gilmer. Membership Inference Attacks against Machine Learning Models.

Physical-World Attacks on Deep Learning Visual Classification

[7] Martín Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, Li Zhang. Deep Learning with Differential Privacy

[8] Diego A. Camacho-Hernández, Victor E. Nieto-Caballero, José E. León-Burguete, Julio A. Freyre-González.Partition Quantitative Assessment (PQA): A quantitative methodology to assess the embedded noise in clustered omics and systems biology data.

[9] Haisheng Su, Jinyuan Feng, Hao Shao, Zhenyu Jiang, Manyuan Zhang, Wei Wu, Yu Liu, Hongsheng Li, Junjie Yan. Complementary Boundary Generator with Scale-Invariant Relation Modeling for Temporal Action Localization: Submission to ActivityNet Challenge 2020.

[10] Fulu Zheng, Yuyu Zhang, Lu Wang, Yadong Wei, Yang Zhao. Engineering Photon Delocalization in a Rabi Dimer with a Dissipative Bath

[11] S. P. Veldanda and P. Domingos. Machine Unlearning: A Survey

[12]. Kevin Eykholt,Ivan Evtimov,Earlence Fernandes,Bo Li,Amir Rahmati, ChaoweiXiao,AtulPrakash,TadayoshiKohno,DawnSong.Robust

[13] Ayush Sekhari, Jayadev Acharya, Gautam Kamath, Ananda Theertha Suresh. Remember What You Want to Forget: Algorithms for Machine Unlearning.

[14] Cynthia Dwork. Differential Privacy

[15] Tom B. Brown, Dandelion Mané, Aurko Roy, Martín Abadi, Justin Gilmer. Towards Deep Learning Models Resistant to Adversarial Attacks

[16] A.Courtoy, Santiago Noguera, Sergio Scopetta.Two-current correlations in the pion in the Nambu and Jona-Lasinio model

[17] James Kirkpatrick, Razvan Pascanu, and Raia Hadsell. Overcoming Catastrophic Forgetting in Neural Networks.

[18] Nicholas Carlini and David Wagner. "Towards Evaluating the Robustness of Neural Networks"

[19] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. MADRY et al. (2017) - Towards Deep Learning Models Resistant to Adversarial Attacks

[20] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Adversarial Patch.