ADScAI
Applied Data Science & Artificial Intelligence Symposium

Model Shifting Unlearning: A Scalable Approach to Data Removal

Lahiruni Chamudika Kumari Pallewela
School Of Computing
APIIT Kandy Campus
Kandy, Sri Lanka
cb010521@students.apiit.lk

***Keywords*—Machine Unlearning, Model-Shifting, Privacy-Preserving AI, Neural Networks, Data Removal**

I. Introduction

In the modern world digital age data has become the main driver of progress for several areas of life. However, the growing dependence upon data increased the alarm over the safety of people's private rights with the advent of regulatory systems such as the European Union's General Data Protection Regulation, which places importance upon the "Right to be Forgotten" [1]. The law grants individuals the right to require the removal of personal details from company databases and presents serious challenges for machine learning (ML) algorithms that are drawing conclusions from huge data sources [2]. Conventional compliance solutions require extensive retraining of machine learning models for the removal of specific data, something that requires much resources. The solution is generally impossible for big models, highlighting the critical need for more efficient solutions. Current practices, such as the use of influence functions and the introduction of noise, attempt to solve the challenge; however, they often sacrifice model performance or are confronted with scalability [3]. Motivated by these challenges our study introduces Model Shifting Unlearning, a novel technique designed to efficiently remove specific data influences from ML models without necessitating full retraining. This method aims to identify and suppress neurons significantly impacted by the data to be forgotten, thereby maintaining overall model integrity and performance. The primary objective of this study is the development of a scalable unlearning framework that preserves model performance and testing of the efficiency of the framework compared with state of the art techniques [4]. With the solution of the great challenge of knowledge removal within machine learning, the study helps advance more ethically accountable and lawfully compliant AI systems.

II. LITERATURE REVIEW

The rapid advancement of machine learning (ML) has raised concerns about data privacy, necessitating machine unlearning techniques that allow models to forget specific information. Machine unlearning ensures that an ML model erases the influence of particular data points, complying with regulations like GDPR and CCPA, which grant individuals the right to data deletion [5]. Traditional unlearning approaches rely on retraining models from scratch after removing data, but this is computationally expensive and impractical [6]. Alternative privacy-preserving techniques, such as differential privacy and federated learning, do not inherently facilitate selective forgetting [7]. More recent unlearning techniques are of two types: exact and approximate. The exact techniques such as SISA (Sharded, Isolated, Sliced, and Aggregated) unlearn just the impacted shard with increased efficiency but are inefficient for complex deep models of learning [8]. The approximate unlearning techniques such as the influence functions and the gradient-based techniques estimate and correct the model's weights but with no complete retraining and with the ability to leave residual traces of the data that compromise the privacy [9]. Early stopping techniques are aimed at preventing data memorization but with no complete removal guaranteed. Existing methods face challenges such as high computational costs, partial retraining requirements, and imperfect data removal. Our proposed Model-Shifting Unlearning technique selectively suppresses neurons responsible for the forget set, efficiently eliminating data influence without full retraining. This approach enhances unlearning accuracy and efficiency, making it more suitable for real-world applications, including regulatory compliance and ethical AI deployment.

III. Materials and Methods

The experiments were conducted using a publicly available medical dataset. The data underwent standard pre-processing steps, including cleaning, normalization and data augmentation, ensuring consistency and reliability.

A. Influence Identification

The first step in the process involves determining the baseline model output by running the trained model over a forget set. The forget set refers to a collection of data that the model should ideally forget or give less importance to in its future predictions. The outputs generated by the model

https://doi.org/10.31705/ADScAI.2025.01

when processing the forget set are considered the reference outputs. These reference outputs are necessary for analyzing the impact of the following interventions. Next, a slight noise is deliberately introduced to the inputs of the forget set. This is done by modifying the inputs in a controlled manner, creating small disturbances. The purpose of this is to analyze how the model responds to these minor changes. The sensitivity analysis is then performed by comparing the model's original outputs and the outputs generated after the noise-induced disturbance. Sensitivity analysis refers to the process of assessing how much a small change in input affects the model's outputs. Specifically, the absolute difference between the two sets of outputs is computed. Neurons that exhibit large changes in output due to the noise are considered highly sensitive to the forget set's influence. The neurons are then ranked by their sensitivity, with the most sensitive neurons considered as crucial contributors to the model's retention of knowledge from the forget set. These neurons are selected for the next phase, where their impact on the model is targeted for suppression.

B. Neuron Suppression

Once the most sensitive neurons are identified, the next step is to suppress their ability to retain the forget set's influence. This is achieved by applying a suppression mask to the weights of these neurons. A suppression mask is a technique used to gradually reduce the contribution of certain neurons to the model's predictions. The suppression mask is implemented using an exponential decay function, which slowly diminishes the weight of the selected neurons over time. This gradual reduction prevents sudden disruptions in the model's ability to learn and generalize. The rationale behind this is that abrupt changes in neuron behavior might impair the model's ability to perform well on other data. By using an exponential decay, the impact of these neurons is reduced over time, allowing the model to learn to ignore the forget set's influence, while still maintaining its ability to generate correct outputs for other data. The mask ensures that this process happens smoothly, so the model's performance is not severely harmed in the process.

C. Fine-Tuning Optimization and Stability

The final phase involves fine-tuning the model to restore lost performance and maintain stability after suppressing certain neurons. This process includes three key steps. First, Gradient-Based Stabilization makes small updates to neurons that are not highly sensitive to the forget set, using model gradients to regain predictive accuracy without relearning forgotten data. Next, Loss Metric-Based Evaluation assesses the model's performance through a loss function, ensuring accuracy remains stable and the unlearning process is effective. Finally, Validation Testing evaluates the model on an independent validation set to confirm it generalizes well to new data without performance degradation.

IV. Results and Discussion

The findings suggest that Model-Shifting Unlearning effectively counteracts the effect of the forget set while retaining a degree of accuracy competitive with that of complete retraining the proposed technique also outperforms competing unlearning techniques with respect to computational efficiency since it demonstrates a shorter running time compared with retraining and with gradient ascent. Unlike the unstable effect of noise injection, Model-Shifting Unlearning enables more structured forgetting by actively reducing the effect of dominant neurons. The framework successfully balances the compromise between processing time and the use of memory. Even though retraining produces the best precision, it calls for much more computing resources. The suggested solution offers a cost-effective alternative that achieves the same extent of unlearning while avoiding the hefty cost of complete retraining. Future work will focus on applying this approach to larger datasets and exploring further optimizations to enhance model stability and generalization.

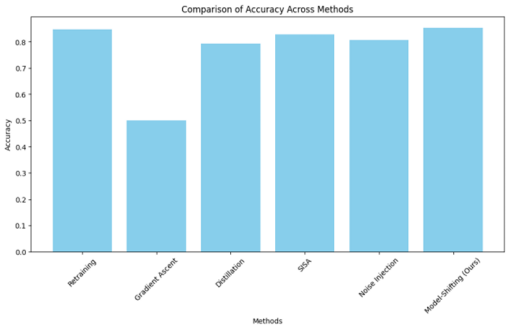

Fig. 1. Image showcasing the accuracy of the machine unlearning methods

TABLE I
TABLE 1: PERFORMANCE COMPARISON OF DIFFERENT MACHINE UNLEARNING METHODS.

Method	Accuracy	Precision	Recall	F1 Score	Time
Retraining	85.42	90.44	81.66	85.82	11.40
Gradient Ascent	87.14	85.85	88.93	87.36	20.58
Distillation	86.82	83.72	91.42	87.40	10.57
SISA	86.64	85.70	87.97	86.82	9.31
Noise Injection	87.45	88.19	86.48	87.33	12.07
Model Shifting	85.35	83.17	88.65	85.82	9.77

References

[1] European Union, "Regulation (EU) 2016/679 of the European Parliament and of the Council," 2016.

[2] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge: Cambridge University Press, 2008.

[3] Varonis, "The Right to Be Forgotten and AI: A Growing Challenge," Varonis Blog, 2024.

[4] J. Bourtoule, V. Chandrasekaran, M. Espig, and others, "Machine Unlearning," arXiv preprint arXiv:2308.07061, 2023.

[5] N. Li et al., "Machine Unlearning: Taxonomy, Metrics, Applications, Challenges, and Prospects," arXiv preprint arXiv:2403.08254, 2024.

[6] W. Wang, Z. Tian, and S. Yu, "Machine Unlearning: A Comprehensive Survey," arXiv preprint arXiv:2405.07406, 2024.

[7] H. Liu, P. Xiong, T. Zhu, and P. S. Yu, "A Survey on Machine Unlearning: Techniques and New Emerged Privacy Risks," arXiv preprint arXiv:2406.06186, 2024.

[8] J. Xu, Z. Wu, C. Wang, and X. Jia, "Machine Unlearning: Solutions and Challenges," arXiv preprint arXiv:2308.07061, 2023.

[9] "Google's Next Big Challenge: Why Should Machine 'Unlearn.'"