Deep Unlearning: Comparative Analysis of Deep Unlearning Techniques on Various Neural Network Models

Ambareesh Ramakrishnan, Gokul Kesavamurthy, Aishwarya Maria Joy
Oregon State University, School of Electrical Engineering and Computer Science,
Kelley Engineering Center, OR, Corvallis, 97331, USA
{ramakria,kesavamg,joyai}@oregonstate.edu

Abstract

Deep learning models often excel at learning complex patterns from data, yet in scenarios requiring compliance with privacy regulations, certain data must be removed from both the dataset and the models without hurting the model accuracy. This study investigates three deep unlearning techniques applied to three neural network architectures: a simple CNN, ResNet-18, and Vision Transformer. Our goal is to assess their effectiveness in facilitating model forgetfulness and compare their performance across diverse model architectures. Through rigorous experimentation, we reveal insights into the efficacy of these techniques, particularly highlighting their effectiveness on the simple CNN model. Furthermore, a cross-comparison with ResNet-18 and Vision Transformer models sheds light on the robustness and adaptability of the unlearning techniques across varying complexities. This analysis not only guides the application of deep unlearning methods but also advances our understanding of model adaptability and flexibility in compliance-driven deep learning contexts.

Moreover, our investigation underscores the importance of developing techniques that not only enhance model performance but also ensure adherence to privacy regulations. By addressing the challenges of data privacy and regulatory compliance within the deep learning paradigm, this research contributes to the development of more responsible and ethically sound AI systems. Additionally, our comparison method is novel, and such comprehensive analysis of unlearning algorithms has not been done before to our knowledge, adding a significant contribution to the existing literature on deep unlearning techniques.

1. Introduction

Deep learning models have revolutionized various fields, demonstrating remarkable capabilities in learning complex patterns from vast amounts of data. Despite the widespread adoption, there's a rising awareness of the need to comply with privacy regulations and address concerns. In many real-world applications, stringent regulatory laws necessitate the removal of certain sensitive data from both the training dataset and the deployed models [12]. Failure to comply with regulations not only poses legal risks but also undermines trust and ethical principles in AI systems. The necessity for deep unlearning is illustrated in Figure 1.

To address these challenges, the concept of "deep unlearning" has emerged as a promising approach to selectively forget specific patterns or information learned by deep learning models [1]. Unlike traditional retraining methods, deep unlearning techniques enable models to adapt to changing conditions or privacy requirements without the need for extensive retraining from scratch. By selectively modifying the parameters or structure of the model, these techniques allow for the removal or suppression of sensitive information while preserving the model's overall performance and functionality.

In this research project, we focus on evaluating the effectiveness of three distinct deep unlearning techniques applied to three different neural network architectures: a simple Convolutional Neural Network (CNN), ResNet-18, and Vision Transformer. Our primary objective is to assess the performance and efficacy of these unlearning algorithms in facilitating model forgetfulness and compliance with privacy regulations. By conducting a comparative analysis across diverse model architectures, we aim to elucidate insights into the relative strengths and limitations of each technique and its adaptability to varying complexities.

The remainder of this paper is organized as follows: Section 2 provides a review of related work in the field of deep unlearning and its applications. Section 3 presents the methodology and experimental setup employed in our study. Section 4 covers experiment results and compares the performance of different deep unlearning techniques across various neural network architectures. Finally, Section 5 concludes the report with a summary of key findings and directions for future research.

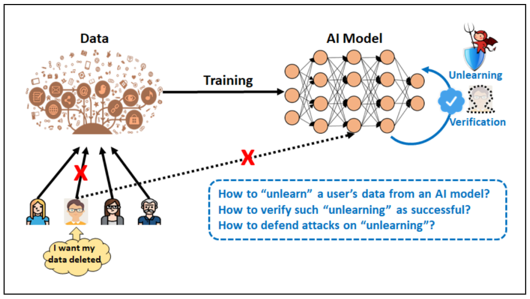

Data
AI Model
Training
Unlearning
Verification
How to "unlearn" a user's data from an AI model?
How to verify such "unlearning" as successful?
How to defend attacks on "unlearning"?
I want my data deleted

Figure 1: Unlearning Scenario

2. Related Work

Recent machine learning advancements have led to diverse techniques for unlearning specific data from trained models, driven by the need to address privacy concerns and comply with regulations such as the "right to be forgotten" under GDPR. Mehta et al.[9] introduce Deep Unlearning via Randomized Conditionally Independent Hessians, employing a variant of the conditional independence coefficient to identify model parameters with semantic overlap, enabling approximate unlearning without inverting large matrices. Meanwhile, Tarun et al.[10] focus on deep regression unlearning, proposing Blindspot unlearning and Gaussian fine-tuning methods tailored for regression problems, showcasing their effectiveness across various applications including computer vision, NLP, and forecasting.

Bourtoule et al. [1] introduce SISA training, a framework designed to expedite unlearning by restricting the influence of specific data points during training, particularly beneficial for stateful algorithms like stochastic gradient descent for deep neural networks. Similarly, Gupta et al. [5] present Adaptive Machine Unlearning, which offers strong provable deletion guarantees for adaptive deletion sequences by leveraging differential privacy and its connection to max information, addressing constraints in handling adaptive unlearning sequences. Further, Tarun et al. [11] propose a Fast Yet Effective Machine Unlearning framework, employing error-maximizing noise generation and impair-repair based weight manipulation to efficiently unlearn single or multiple classes of data from deep networks without necessitating access to the full training data.

Chen et al. [2] propose Boundary Unlearning, a rapid and effective technique that shifts the decision boundary of deep neural network models to facilitate unlearning specific classes of data, demonstrating significant speed-ups compared to retraining from scratch. Meanwhile, Lee and Woo [7] introduce UNDO, a two-step unlearning method that selectively disrupts and repairs decision boundaries at both coarse-grained and fine-grained levels, achieving effective and accurate unlearning while preserving overall classification performance. Additionally, Kim and Woo [6] propose an Efficient Two-Stage Model Retraining approach for machine unlearning in computer vision classification models, leveraging contrastive labels and knowledge distillation to efficiently remove requested datasets from trained models.

Marchant et al. [8] underscore the vulnerability of certified machine unlearning to poisoning attacks, revealing how attackers can exploit the computational cost of data removal to trigger complete retraining through strategically designed training data. Finally, Zhang et al. [12] provide a comprehensive review of machine unlearning approaches, discussing their applications, privacy concerns, and future research challenges, offering insights into the evolving landscape of privacy-preserving methods in machine learning.

These techniques provide diverse approaches to efficiently unlearn specific data from trained models, each with its own advantages, limitations, and considerations for practical implementation in privacy-sensitive applications.

This study assesses three deep unlearning techniques [11, 4, 3] on CNN, ResNet-18, and Vision Transformer models, aiming to evaluate their effectiveness in facilitating model forgetfulness. Through experimentation, insights into efficacy are revealed, emphasizing the techniques' effectiveness on the simple CNN model. A cross-comparison with ResNet-18 and Vision Transformer models highlights robustness and adaptability. This analysis advances our understanding of model adaptability and flexibility in compliance-driven deep learning.

3. Technical Approaches

In this study, we selected three algorithms based on a comprehensive literature review. These algorithms were chosen for their relevance to the task of deep unlearning and their potential effectiveness across different neural network architectures. In the following sections, we will delve into each of these selected approaches, providing detailed explanations of their underlying principles, implementation methodologies, and comparative analyses of their performance on our chosen neural network models.

3.1. Amnesiac Unlearning

Amnesiac unlearning is a method proposed by the authors to selectively undo the learning steps that involved sensitive data in a trained neural network model. This approach aims to efficiently remove the learned information about sensitive data from the model while preserving the overall model performance on non-sensitive data.

3.1.1 Method Description

During the training process, the model owner keeps a record of which training examples appeared in which batches, as well as the parameter updates from each batch. When a data removal request comes for specific sensitive data, the model owner can undo the parameter updates from only the batches containing that sensitive data.

Formally, let's consider a model M trained for E epochs, each consisting of B batches, with the initial model parameters θ_{initial}. The learned model parameters θ_M can be expressed as:

$$\theta_M = \theta_{\text{initial}} + \sum_{e=1}^{E} \sum_{b=1}^{B} \Delta\theta_{e,b} \tag{1}$$

Where $\Delta\theta_{e,b}$ is the parameter update from batch b in epoch e.

During training, the model owner keeps a list SB of which batches contained the sensitive data. They also maintain the model parameter updates from each batch $sb \in SB$.

After training, a protected model M' can be produced using amnesiac unlearning by removing the parameter updates from each batch $sb \in SB$ from the learned parameters θ_M. This can be observed in the following equation (2).

$$\theta_{M'} = \theta_M - \sum_{sb=1}^{SB} \Delta\theta_{sb} \tag{2}$$

3.1.2 Advantages and Considerations

Amnesiac unlearning offers precise data removal capabilities, allowing targeted elimination of specific data segments, such as individual examples or sets of examples, while minimizing disruption to the overall model. This feature is particularly beneficial in scenarios requiring protection of individual privacy within the dataset. However, a potential drawback lies in the storage overhead incurred by tracking parameter updates from each batch, especially for large-scale models. Additionally, the efficacy of the unlearning process may vary based on the number of batches processed, with lower batch counts resulting in minimal impact on model performance. Nonetheless, amnesiac unlearning remains an efficient and practical solution for removing sensitive data from trained neural network models, ensuring compliance with regulations such as the GDPR's "right to be forgotten" while preserving model integrity.

3.2. Knowledge Transfer Based Unlearning

3.2.1 Method Description

This deep unlearning method employs a teacher-student framework with competent and incompetent teachers to induce forgetting in deep neural networks. The aim is to remove information linked to a specific set of data samples (forget set) from a pretrained model without retraining from scratch. With the increasing importance of data privacy regulations granting individuals the right to be forgotten, machine unlearning has become a crucial aspect of model maintenance.

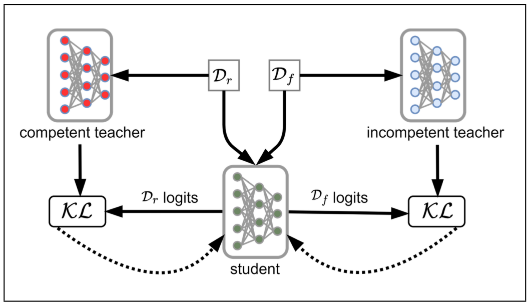


Figure 2: Knowledge Transfer Student-Teacher Framework

This method utilizes a competent teacher (T_s), an incompetent teacher (T_d), and a student model (S). The competent teacher represents the fully trained model, while the incompetent teacher is randomly initialized. The student model is initialized with the parameters of the competent teacher.

The dataset is modified in such a way that each data point is assigned an unlearning label, l_u, which is 1 if the sample belongs to D_f (forget dataset) and 0 if it belongs to D_r (retain dataset). The subset used for unlearning is $\{(x_i, l_{ui})\}_{i=1}^{p}$, where p is the total number of samples, and l_{ui} is the unlearning label corresponding to each sample x_i.

Now, the unlearning objective is formulated as:

$$L(x, l_u) = (1-l_u)*\text{KL}(T_s(x)||S(x))+l_u*(\text{KL}(T_d(x)||S(x))) \tag{3}$$

where l_u is the unlearning label (1 for forget set, 0 for retain set), and KL denotes the Kullback-Leibler divergence between the output distributions of the teachers and the student. The student is trained to minimize this loss, selectively transferring knowledge from the incompetent teacher for the forget set and from the competent teacher for the retain set. This framework is illustrated in Figure 2. The key concept involves selectively transferring knowledge from the competent and incompetent teachers to the student model. For the forget set (D_f), the student learns to mimic the incompetent teacher (T_d), which possesses random knowledge about the forget samples. This facilitates the removal of information exclusively pertaining to those samples from the student model.

For the retain set (D_r), the student learns from the competent teacher (T_s), which holds accurate knowledge about these samples. This ensures the preservation of information related to the retain set in the student model. Selective knowledge transfer is achieved by minimizing the KL divergence between the student and the respective teacher (competent or incompetent) for each sample, depending on its unlearning label (l_u).

3.2.2 Advantages and Considerations

The method involves mimicking the incompetent teacher's predictions for the forget set, introducing controlled randomness while preserving a degree of generalization. However, this approach risks contaminating the retain set's information with the teacher's random knowledge. To address this, the method selectively integrates accurate knowledge from a competent teacher, minimizing KL divergence to ensure retention of correct knowledge about the retain set.

This technique supports various unlearning scenarios, including single-class, multiple-class, sub-class, and random subset forgetting, achieved by setting appropriate unlearning labels (l_u) without altering the method itself. Unlike existing methods with stringent training constraints or the need for additional models, this approach is adaptable to any pretrained model without prior training knowledge, ensuring computational efficiency without requiring model retraining or additional training phases.

3.3. Impair-Repair Based Unlearning

3.3.1 Method Description

In the context of deep networks, this approach addresses the unlearning problem where a complete training dataset D_c consists of n samples and K total classes, denoted as $D_c = \{(x_i, y_i)\}_{i=1}^{n}$, where $x \in X \subset \mathbb{R}^d$ represents the inputs and $y \in Y = \{1, \ldots, K\}$ are the corresponding class labels. Given forget and retain classes Y_f and Y_r, respectively, with $D_f \cup D_r = D_c$ and $D_f \cap D_r = \varnothing$, the objective is to derive a new set of weights θ_{D_r} using the trained model f and a subset of retain images $D_r^{sub} \subset D_r$ that excludes information regarding D_f, similarly to a model that has not encountered D_f in the parameter and output space.

To achieve unlearning, this approach first learns a noise matrix N for each class in Y_f using the trained model. Then the model modifies to fail in classifying samples from the forget set D_f while maintaining accuracy for samples from the retain set D_r. This is accomplished by using a small subset of samples D_r^{sub} drawn from the retain dataset D_r.

This approach focuses on learning an error-maximizing noise matrix for the unlearning class to overwrite previously learned network weights and induce unlearning. The goal is to create a correlation between the noise matrix N and the unlearning class label, $f : N \rightarrow Y_f$, where $N \neq X$. By minimizing the loss function (4),

$$\arg\min_{N} E(\theta)[-L(f, y) + \lambda||w_{\text{noise}}||], \tag{4}$$

it optimizes the noise matrix to maximize the model's classification loss while regularizing to prevent overfitting. The regularization term $\lambda\|w_{noise}\|$ manages the trade-off between the two objectives, ensuring that the noise values do not become excessively large. The noise matrix is learned separately for each class of data, facilitating efficient unlearning in deep networks without extensive computational overhead. Finally, the unlearning with Impair-Repair algorithm combines the noise matrix with the retain data subset D_r^{sub}, training the model for one epoch to induce unlearning (impair) and another epoch to restore accuracy (repair). This approach ensures effective unlearning while minimizing disruption to the model's performance. Figure 3 provides a detailed description of this framework.

3.3.2 Advantages and Considerations

This method excels in efficiency, surpassing traditional approaches like retraining and Fisher forgetting in computational speed. Optimizing the error-maximizing noise matrix and the impair-repair mechanism allows swift unlearning without heavy computational resources. Furthermore, its scalability enables it to unlearn multiple classes of data with minimal computational overhead, rendering it suitable for large-scale problems and intricate models.

However, certain limitations merit consideration. The method's scope primarily encompasses targeted class unlearning based on error-maximizing noise generation, limiting its applicability to scenarios involving random sample unlearning. Additionally, a trade-off exists between efficiency and accuracy, with a slight potential decrease in overall model accuracy, particularly for targeted classes. This trade-off necessitates careful consideration based on the specific application requirements. Nonetheless, this method represents a significant advancement in deep unlearning, offering a rapid and scalable solution for removing specific classes of data from deep neural network models. Its efficacy and efficiency contribute significantly to addressing privacy concerns and ensuring model compliance across diverse machine learning applications.

4. Experiment and Results

4.1. Evaluation Metric

Before delving into the experimental details and their outcomes, it's essential to establish an evaluation metric to effectively compare the performance of various algorithms. In this context, we introduce a metric termed "**forget quality**" to assess the efficacy of these algorithms. Forget quality quantifies how well an unlearned model performs compared to a model trained solely on the retain dataset, often referred to as the Gold Model.

To measure forget quality, we rely on comparing the performance metrics of the unlearned model with those of the Gold Model. For convenience and clarity, we define the following terms:

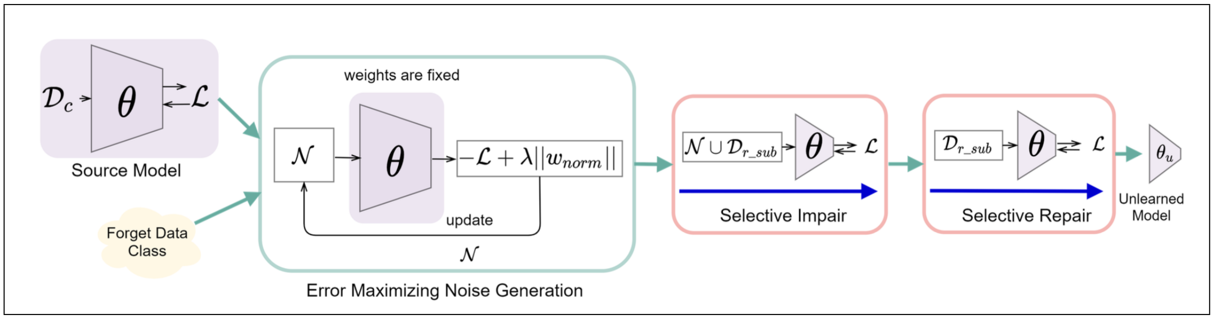


Figure 3: Impair-Repair Architecture

$$\begin{aligned} RA^U &= RA(\{U_1, \ldots, U_N\}), \\ TA^U &= TA(\{U_1, \ldots, U_N\}), \\ RA^R &= RA(\{R_1, \ldots, R_N\}), \\ TA^R &= TA(\{R_1, \ldots, R_N\}), \end{aligned} \tag{5}$$

where RA^U and TA^U represent the retain and test accuracy, respectively, of unlearned models. Similarly, RA^R and TA^R denote the retain and test accuracy, respectively, for the ideal unlearning algorithm of retraining. A key criterion for assessing an unlearning algorithm's effectiveness is its ability to maintain RA^U and TA^U values close to their RA^R and TA^R counterparts.

Forget quality (F) is then defined as the product of the ratio of retain accuracy to test accuracy for both unlearned and retrained models:

$$F = \frac{RA^U}{RA^R} \times \frac{TA^U}{TA^R}. \tag{6}$$

The value of forget quality serves as a measure of how closely the unlearned model resembles the Gold Model. Ideally, an unlearned model with a forget quality value of 1 signifies optimal performance, indicating minimal degradation in accuracy compared to the Gold Model.

4.2. Experiment Setup

In our experimental setup, we investigated the efficacy of three different deep unlearning approaches across various neural network architectures and datasets. We began by examining the performance of these approaches on a basic Convolutional Neural Network (CNN) built from scratch. This CNN consisted of multiple layers with 10 output neurons and was trained to classify handwritten digits ranging from 0 to 9 using the MNIST dataset. For this experiment, we designated one digit as the forget set and retained the remaining nine digits for evaluation.

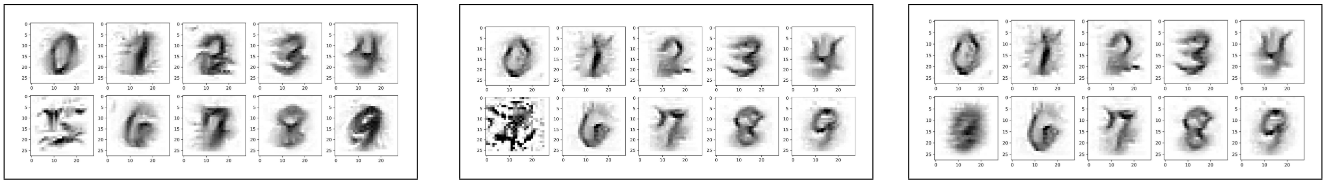
We then focused on the ResNet18 model, recognized for its deep layer structure and skip connections. We adapted the ResNet18 backbone by modifying its fully connected and output layers and retrained it to classify objects in the CIFAR-20 dataset. Here, we selected one class from CIFAR-20 as the forget set and retained the rest of the dataset for training and evaluation purposes.

Following the ResNet18 experiment, we proceeded to evaluate the deep unlearning approaches on the Vision Transformer (ViT) model. The Vision Transformer is a state-of-the-art architecture that utilizes self-attention mechanisms for image classification tasks. We utilized pre-trained ViT weights and fine-tuned the model's output layer to classify images in the CIFAR-20 dataset. Similar to the ResNet18 experiment, we designated one class from CIFAR-20 as the forget set and retained the remaining classes for training and evaluation.

Our experimental setup aimed to provide comprehensive insights into the deep unlearning approaches across various model complexities and datasets. We began with a simple CNN and gradually transitioning to more sophisticated architectures like ResNet18 and Vision Transformer, to assess the adaptability and robustness of the unlearning methods. Exploring unlearning on Vision Transformer posed a significant challenge due to its strong generalizability, adding intrigue to our study. Through these experiments, we aimed to gain valuable insights into the effectiveness and limitations of deep unlearning in various real-world scenarios.

4.3. Results

Table 1 provides a comparative analysis of forget quality across CNN, ResNet18, and ViT models. The forget quality percentages indicate the effectiveness of each approach in removing specific data from the models. Meanwhile, Table 2 presents a comparison of forget and retain dataset accuracies for the Amnesiac Unlearning algorithm across the same three models. This comparison illustrates how well the algorithm performs in forgetting targeted data while retaining the accuracy on the remaining dataset. Both tables offer valuable insights into the performance of unlearning techniques across different neural network architectures.



(a) Amnesiac Unlearned Model: Unlearning Digit 5

(b) Knowledge Transfer-Based Unlearned Model: Unlearning Digit 5

(c) Impair-Repair-Based Unlearned Model: Unlearning Digit 5

Figure 4: The results of the Model Inversion Attack are presented for all three unlearning approaches applied to the CNN model trained on the MNIST digits dataset. Specifically, the data corresponding to digit 5 is targeted for unlearning using these approaches.

Unlearning Approach	**CNN**	**ResNet18**	**ViT**
Amnesiac Unlearning	68%	72%	35%
Knowledge Transfer	77%	88%	52%
Impair-Repair	89%	83%	49%

Table 1: Comparative Analysis of **Forget Quality**

Amnesiac Unlearning	**CNN**	**ResNet18**	**ViT**
Forget Accuracy	38%	35%	71%
Retain Accuracy	81%	80%	79%

Table 2: Forget and Retain dataset accuracy comparison of **Amnesiac Unlearning algorithm**

4.3.1 Unlearning on CNN with MNIST Digits Dataset

First we tested all the three approaches with MNIST Digits dataset with a simple CNN neural network architecture. We calculate the forget quality for the unlearned models trained using the three approaches. To get a better visualization of the unlearned model, we perform model inversion attack only on the CNN model trained on MNIST Digits dataset.

Model inversion attack is a privacy threat in machine learning where an adversary attempts to reverse-engineer or infer sensitive information about the training data from a deployed model. The adversary uses the model's output to reconstruct the original inputs used during training, exploiting predictions/gradients, or other outputs to recover private or sensitive information such as images or other data points.

Based on the outcomes of the model inversion attack, it is evident that in Amnesiac unlearning, residual traces of digit 5 remain, unlike in Knowledge Transfer and Impair-Repair-based unlearning methods, the models exhibit greater resilience in obliterating the presence of digit 5 from their representations. This is substantiated by the results, indicating these models fail to reproduce digit 5 since the majority of its feature information has been eradicated from the model's learned representations (Figure 4). Furthermore, the forget quality for all the approaches is summarized in Table 1.

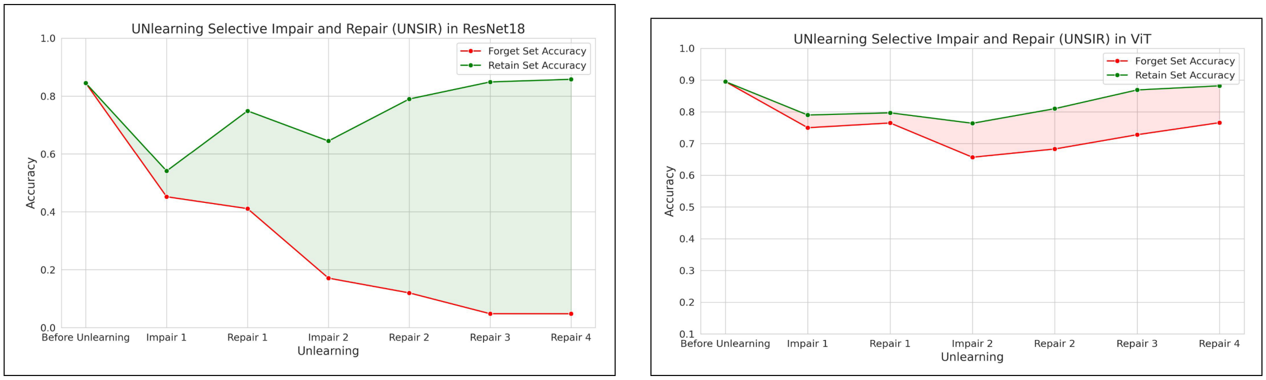
4.3.2 Unlearning on ResNet18 with CIFAR-20 Dataset

The Amnesiac unlearning approach consistently exhibits a slight lag behind the other two methodologies, as evidenced by the final forget and retain accuracy metrics in Table 2. However, the Knowledge Transfer-based and Impair-Repair methods show comparable efficacy across various scenarios. Notably, the Impair step within the Impair-Repair framework tends to lower overall model accuracy, particularly affecting forget set accuracy more than the retain set accuracy. Conversely, the Repair step significantly boosts retain set accuracy through training with the retain set, but at the expense of notable reductions in forget set accuracy.

With successive Impair steps, forget set accuracy experiences a significant decline, while retain set accuracy notably improves after the final three Repair steps. This trend underscores the robust performance and high-quality forget/unlearn capability of the Impair-Repair approach. A similar pattern is discernible in the Knowledge Transfer-based algorithm. Here, the accuracy tends to improve as the loss function incorporating KL divergence converges over the course of training. The accuracies observed during the unlearning process are presented in Figure 5, 6, 7 and 8.

4.3.3 Unlearning on Vision Transformer (ViT) Model using CIFAR-20 Dataset

Upon comprehensive analysis, it becomes evident that the performance of the Amnesiac unlearning approach consistently exhibits a slight lag behind the other two methodologies. This trend is observable in Table 2, where it is apparent that the Amnesiac approach demonstrates relatively lower efficacy in forgetting specific data points. This observation is further substantiated by examining the final forget and retain accuracy metrics produced by the algorithms. Comparing the Knowledge Transfer based and Impair-Repair methods, both show similar efficacy. In the Impair-Repair approach, the impair step diminishes overall model accuracy, notably affecting forget set accuracy more than retain set accuracy. The repair step enhances retain set



(a) Unlearning on ResNet18

(b) Unlearning on ViT

Figure 5: Forget and Retain Data Accuracy of **Impair-Repair Unlearning Approach** on ResNet18 and ViT model

Epoch	Forget Acc.	Retain Acc.
Before Unlearning	84.55%	84.52%
Impair Epoch 1	45.2%	54.1%
Repair Epoch 1	41.1%	74.9%
Impair Epoch 2	17.1%	64.5%
Repair Epoch 2	12.02%	79.07%
Repair Epoch 3	4.8%	84.9%
Repair Epoch 4	4.78%	85.8%

(a) Unlearning on ResNet18

Epoch	Forget Acc.	Retain Acc.
Before Unlearning	89.52%	89.49%
Impair Epoch 1	75.4%	75.2%
Repair Epoch 1	76.5%	79.7%
Impair Epoch 2	65.7%	76.4%
Repair Epoch 2	68.31%	81.01%
Repair Epoch 3	72.8%	86.91%
Repair Epoch 4	76.6%	88.2%

(b) Unlearning on ViT

Figure 6: Accuracy Comparison of **Impair-Repair Unlearning Approach** on ResNet18 and ViT Model

accuracy by utilizing it for training. Notably, forget set accuracy also improves, albeit to a lesser extent compared to the CNN and ResNet18 models, due to the Vision Transformer's attention mechanism generalizability. Despite all impair steps, forget set accuracy remains relatively high. However, after two repair steps, retain set accuracy indicates strong model performance but with a notable deficiency in forget/unlearn quality.

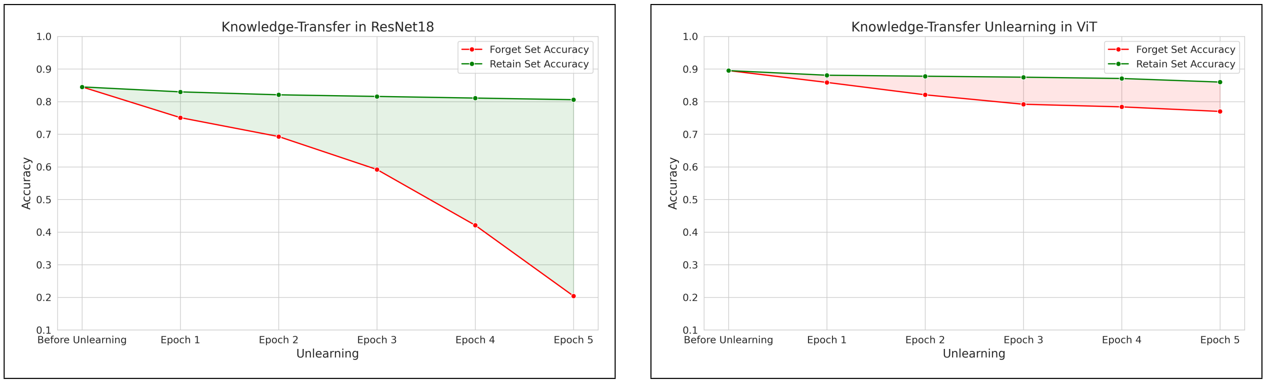
Similarly, there exists a slight difference between retain and forget accuracy for every epoch in knowledge-transfer based unlearning. The model doesn't forget the data very well compared to CNN and ResNet18. This is due to the highly generalizable nature of the ViT model. The accuracies observed during the unlearning process are presented in Figure 5, 6, 7 and 8.

5. Conclusion and Future Work

In conclusion, our study presents a thorough comparative analysis of three distinct unlearning methodologies—Amnesiac Unlearning, Knowledge Transfer-based Unlearning, and Impair-Repair-based Unlearning—across various neural network models and datasets. Through comprehensive experimentation, we unveil nuanced performance differences among these approaches, particularly influenced by the size and characteristics of the forget dataset. While Amnesiac Unlearning shows promise in scenarios with minimal data volumes, it encounters scalability challenges with larger forget datasets. In contrast, Knowledge Transfer-based Unlearning and Impair-Repair-based Unlearning exhibit consistent and robust performance across diverse experimental scenarios, positioning them as compelling choices for broader unlearning applications.

Our research sets the stage for future exploration of tailored unlearning techniques for emerging architectures like Generative Adversarial Networks (GANs) and Transformer-based models. By delving into these advanced frameworks, we seek fresh insights into unlearning algorithms' adaptability and effectiveness. Our forthcoming efforts will focus on thorough comparative analyses to uncover the strengths and weaknesses of these evolving methodologies, contributing significantly to the ongoing discussion in machine learning and artificial intelligence.

Moreover, our unique comparative analysis methodology represents a novel approach, as comprehensive analysis of unlearning algorithms in different architectures with diverse complexities has not been previously under-



(a) Unlearning on ResNet18

(b) Unlearning on ViT

Figure 7: Forget and Retain Data Accuracy of **Knowledge-Transfer Unlearning Approach** on ResNet18 and ViT model

Epoch	Forget Acc.	Retain Acc.
Before Unlearning	84.55%	84.52%
Epoch 1	75.1%	83.0%
Epoch 2	69.3%	82.1%
Epoch 3	59.2%	81.6%
Epoch 4	42.1%	81.1%
Epoch 5	20.4%	80.6%

(a) ResNet18 Model

Epoch	Forget Acc.	Retain Acc.
Before Unlearning	89.52%	89.49%
Epoch 1	85.9%	88.1%
Epoch 2	82.1%	87.8%
Epoch 3	79.2%	87.5%
Epoch 4	78.4%	87.3%
Epoch 5	77.0%	86.1%

(b) ViT Model

Figure 8: Accuracy Comparison of **Knowledge-Transfer Unlearning Approach** on ResNet18 and ViT Model

taken to our knowledge. This significant contribution enriches the existing literature on deep unlearning techniques, improving our grasp of different methodologies' strengths and weaknesses, guiding future research and practical applications in privacy-sensitive domains.

References

[1] L. Bourtoule, V. Chandrasekaran, C. A. Choquette-Choo, H. Jia, A. Travers, B. Zhang, D. Lie, and N. Papernot. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy (SP)*, pages 141–159. IEEE, 2021.

[2] M. Chen, W. Gao, G. Liu, K. Peng, and C. Wang. Boundary unlearning: Rapid forgetting of deep networks via shifting the decision boundary. In *Proceedings of the IEEE/CVF Conf. on CV and Pattern Recognition*, pages 7766–7775, 2023.

[3] V. S. Chundawat, A. K. Tarun, M. Mandal, and M. Kankanhalli. Can bad teaching induce forgetting? unlearning in deep networks using an incompetent teacher. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 7210–7217, 2023.

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

[5] V. Gupta, C. Jung, S. Neel, A. Roth, S. Sharifi-Malvajerdi, and C. Waites. Adaptive machine unlearning. *In Neural Information Processing Systems*, 34:16319–16330, 2021.

[6] J. Kim and S. S. Woo. Efficient two-stage model retraining for machine unlearning. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition*, pages 4361–4369, 2022.

[7] S. Lee and S. S. Woo. Undo: Effective and accurate unlearning method for deep neural networks. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 4043–4047, 2023.

[8] N. G. Marchant, B. I. Rubinstein, and S. Alfeld. Hard to forget: Poisoning attacks on certified machine unlearning. In *Proceedings of the AAAI Conf. on AI*, volume 36, pages 7691–7700, 2022.

[9] R. Mehta, S. Pal, V. Singh, and S. N. Ravi. Deep unlearning via randomized conditionally independent hessians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10422–10431, 2022.

[10] A. K. Tarun, V. S. Chundawat, M. Mandal, and M. Kankanhalli. Deep regression unlearning. In *International Conference on ML*, pages 33921–33939. PMLR, 2023.

[11] A. K. Tarun, V. S. Chundawat, M. Mandal, and M. Kankanhalli. Fast yet effective machine unlearning. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[12] H. Zhang, T. Nakamura, T. Isohara, and K. Sakurai. A review on machine unlearning. *SN Comp. Science*, 4(4):337, 2023.