SALSA: A Secure, Adaptive and Label-Agnostic Scalable Algorithm for Machine Unlearning

Owais Makroo[*1] **Atif Hassan**[*2] **Swanand Khare**[1]

[1]Department of Mathematics, IIT Kharagpur, Kharagpur, West Bengal, India
[2]Department of Artificial Intelligence, IIT Kharagpur, Kharagpur, West Bengal, India

Abstract

Machine Learning as a Service (MLaaS) has simplified access to powerful machine learning models but faces challenges in complying with the "right to be forgotten" while resisting adversarial threats. Machine Unlearning (MU) addresses these issues by enabling selective data removal from models. However, existing methods are slow, label-dependent, vulnerable to black-box attacks, and computationally impractical for large-scale MLaaS deployments. We introduce SALSA, a Secure, Adaptive, Label-Agnostic, Scalable Algorithm for efficient and robust machine unlearning tailored to classification tasks in MLaaS. SALSA redistributes the class-wise predicted probabilities of data to be forgotten and optimizes a novel loss function that minimizes the divergence between redistributed and predicted probabilities while anchoring model parameters near their initialization. This ensures simultaneous unlearning and generalization. SALSA requires neither labels nor access to the remaining data, making it ideal for MLaaS environments. It is exceptionally fast, achieving at least $25\times$ faster unlearning, on average, than the fastest baseline, while consistently outperforming five state-of-the-art MU techniques across eight metrics on benchmark datasets. Experiments on synthetic data show that SALSA's altered decision boundaries closely approximate exact unlearning. Rigorous evaluations against state-of-the-art black-box attacks demonstrate its resilience to security threats. Thus, SALSA redefines practical machine unlearning, offering a scalable and resilient solution for safeguarding privacy in modern MLaaS systems.

[*]These authors contributed equally to this work

1 INTRODUCTION

Machine learning (ML), particularly deep learning (DL), has revolutionized data-driven services, achieving remarkable performance in domains such as computer vision [He et al., 2016, Dosovitskiy et al., 2021], natural language processing [Brown et al., 2020b, OpenAI, 2023], and speech recognition [Radford et al., 2023, Baevski et al., 2020]. At the core of this progress lies the emergence of deep neural network foundation models that leverage billions of parameters to deliver exceptional performance across diverse tasks Xi et al. [2023]. ML's success has catalyzed the widespread adoption of cloud-based platforms known as Machine Learning as a Service (MLaaS) that democratize access to powerful predictive and analytic tools by allowing users to train, fine-tune, and deploy models without managing complex computational infrastructure. These platforms abstract the complexities of machine learning through APIs, offering benefits such as scalability, cost-effectiveness, and enhanced privacy by separating user data from service providers during deployment [Shmueli et al., 2023].

However, MLaaS is not without its own security and vulnerability issues. Models trained on sensitive data are susceptible to memorizing and exposing private information [Wu et al., 2022, Carlini et al., 2023]. Such vulnerabilities are particularly critical in cloud platforms, where deployed models interact with potentially malicious users, risking data leakage and exploitation through attacks like membership inference and model inversion [Hu et al., 2024b, Shokri et al., 2017a]. In response, government regulations mandate the "right to be forgotten," requiring the effective removal of personal data upon request. While straightforward in storage systems, enforcing this in trained ML models remains a formidable challenge [Thudi et al., 2022b]. Machine Unlearning (MU) has emerged as a potential solution, aiming to erase the influence of specific data points while preserving model performance [Xu et al., 2024, Bourtoule et al., 2021].

1.1 MOTIVATION

A naïve approach to machine unlearning (MU) involves retraining the model from scratch on the remaining data to guarantee complete unlearning. While effective, this approach is computationally prohibitive for modern deep learning models and impractical in Machine Learning as a Service (MLaaS) settings, where servers hosting models lack access to the original training data. To address these challenges, researchers have proposed efficient MU techniques categorized as exact [Bourtoule et al., 2021, Yan et al., 2022] and approximate unlearning [Tarun et al., 2024, Wang et al., 2023, Thudi et al., 2022a]. However, most methods rely on access to remaining training data or labels, which is often unavailable in MLaaS environments due to privacy constraints [Shen et al., 2024]. Exact methods, such as influence functions, require computationally expensive Hessian inversions [Chen et al., 2023, Warnecke et al., 2023], while approximate methods [Huang et al., 2024, Tarun et al., 2024] involve iterative fine-tuning, further limiting scalability.

MU techniques must also address adversarial threats in MLaaS. Models are vulnerable to membership inference Ding et al. [2025], model inversion [Hu et al., 2024b], and malicious unlearning, where attackers exploit unlearning requests to cause over-unlearning, degrading model utility Hu et al. [2024a]. These risks are particularly severe in classification tasks central to MLaaS applications, such as facial recognition [Nair et al., 2023], anomaly detection [Du et al., 2019], and medical diagnosis [Zhou et al., 2023], where breaches or malicious unlearning can have dire consequences. Addressing these computational and security challenges is critical for advancing practical and resilient machine unlearning solutions.

1.2 OUR CONTRIBUTION

We propose SALSA, a Secure, Adaptive and Label-agnostic Scalable Algorithm for Machine Unlearning specifically designed for classification tasks. We introduce a new strategy to redistribute the class-wise predicted probabilities of a model for a given set of samples that need to be forgotten. A novel loss function is then employed to implement unlearning while maintaining generalization performance. This is achieved by simultaneously minimizing the divergence between the redistributed and predicted class-wise probabilities as well as the Euclidean distance between the original and current model parameters. Iteratively fine-tuning a pretrained model using this process results in a computationally efficient unlearning approach that we empirically find to converge within a few steps. Thus, unlike prior techniques, SALSA solely relies on samples that need to be forgotten without requiring corresponding label information to perform unlearning while preserving the model's generalization on the remaining data without its explicit utilization.

We extensively evaluate SALSA under diverse settings. First, we observe that on non-linear synthetic datasets, across multiple unlearning paradigms (sample-wise, subclass-wise, and class-wise unlearning), the altered decision boundary of our unlearned models closely approximate those of exact unlearning. Second, we evaluate SALSA's efficacy on three benchmark datasets, CIFAR10, SVHN and TinyImageNet using ResNet18 and Swin Transformer models, under different unlearning paradigms (sample-wise and class-wise unlearning) against five state-of-the-art machine unlearning methods. SALSA consistently outperforms all considered baselines under eight different evaluation metrics across all datasets while being at least $25\times$ faster, on average, than the fastest baseline. Finally, to ensure the security of the unlearning process, we test SALSA against three prominent black-box attacks that are possible in the MLaaS setting, namely, membership inference, model inversion, and malicious unlearning (over-unlearning). The attacks consistently fail to recover any information about samples unlearned using our approach while also remaining unsuccessful in compromising the performance of the end model. Thus, our proposed algorithm effectively mitigates threats encountered by MLaaS while maintaining high performance.

2 RELATED WORK

2.1 MACHINE UNLEARNING

Introduced by Cao and Yang [2015], machine unlearning focuses on removing specific data influences from trained models. While retraining without the samples to be forgotten ensures complete unlearning, its computational cost is prohibitive for large-scale models like GPT-3 [Brown et al., 2020a], which requires 34 days on 1024 GPUs for retraining [Narayanan et al., 2021]. To address this, efficient unlearning strategies have emerged, categorized as exact and approximate.

Exact Unlearning: Exact unlearning methods, such as SISA [Bourtoule et al., 2021], retrain only on affected data shards while DaRE [Brophy and Lowd, 2021] selectively retrains parts of random forests, reducing overhead but requiring access to the training dataset, an issue in privacy-centric MLaaS environments.

Approximate Unlearning: Approximate methods adjust model parameters without full retraining. Amnesiac ML [Graves et al., 2021] removes gradient updates corresponding to the samples that need to be forgotten to achieve unlearning but risks residual influence. Influence-based approaches [Guo et al., 2020, Izzo et al., 2021, Chen et al., 2023, Warnecke et al., 2023] use influence functions but face scalability issues due to costly inverse Hessian computations. Gradient-based techniques offer practical alternatives. Warnecke et al. [2023] overwrite unlearned data contributions, while methods like SalUn [Fan et al., 2024], SFTC [Perifa-

nis et al., 2024], and LAF [Shen et al., 2024] refine model weights or rely on biased labeling strategies for efficiency. Recent works, including FEMU [Tarun et al., 2024] and SFRon [Huang et al., 2024] achieve scalable and practical unlearning, aligning well with MLaaS requirements.

2.2 BLACK-BOX THREATS IN MLAAS

Black-box attacks present significant challenges to privacy and security in MLaaS.

Membership Inference Attacks (MIAs): MIAs determine training data membership by exploiting overfitting patterns in model outputs [Shokri et al., 2017b]. Advanced methods infer unlearned data membership by analyzing confidence vectors before and after unlearning [Hu et al., 2024c, Chen et al., 2021, Lu et al., 2022, Gao et al., 2022], with top-1 confidence scores enhancing efficacy [Lu et al., 2022].

Model Inversion: Model inversion reconstructs training data from outputs, transitioning from white-box Fredrikson et al. [2014, 2015] to black-box settings. Approaches like LBMI [Yang et al., 2019] leverage autoencoders, while MIRROR [An et al., 2022] and BREP-MI [Kahla et al., 2022] exploit residual data influence using GANs and hard-label outputs, respectively.

Malicious Unlearning: Malicious unlearning degrades model performance during the unlearning phase. Hu et al. [2024a] showed that by pushing data closer to the decision boundary, over-unlearning increases misclassification risks. This threat is amplified in black-box MLaaS settings, where limited transparency exposes models to exploitation.

3 PRELIMINARY

Notations: Let $\mathcal{D} = \{z_i\}_{i=1}^{n}$ be a dataset containing n data points where each samples is $z_i = (x_i, y_i)$. Here, $x_i \in \mathbb{R}^d \sim \mathcal{P}$ is a feature vector assumed to be sampled from an underlying distribution $\mathcal{P}$ while $y_i \in \{1, \cdots, c\}$ is the target/label and c is the number of classes. Let $\mathcal{D} = \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{test}}$ where $\mathcal{D}_{\text{train}}$ is the train set and $\mathcal{D}_{\text{test}}$ is the test set used for model training and evaluation, respectively. Let $\mathcal{D}_u = \{z_i\}_{i=1}^{n_u} \subset \mathcal{D}_{\text{train}}$, denote a subset of training samples to be unlearned, termed the *forget set*. Here, n_u is the number of samples to be unlearned. The remaining samples, termed *retain set* are denoted as, $\mathcal{D}_r = \mathcal{D}_{\text{train}} \setminus \mathcal{D}_u = \{z_i\}_{i=1}^{n_r}$ where $n_r = n - n_u$. Let a machine learning model, trained on $\mathcal{D}_{\text{train}}$, referred to as the original pre-trained model, be $f_\theta : \mathbb{R}^d \to \mathbb{R}^c$ parameterized by $\theta = [\theta_1, \cdots, \theta_L]$ where L is the depth/number of layers of the model. Retraining the original model from scratch on $\mathcal{D}_r$ yields f_{θ_*} which is considered as the oracle for unlearning.

Evaluation Metrics: We assess unlearning using eight metrics: forgetting and retain accuracies on the train set (FA_{tr} and RA_{tr}, respectively) and test sets (FA_{te} and RA_{te}, respectively) to measure unlearning and generalization performance, test accuracy (TA) for overall generalization, average discrepancy (Avg. D), defined as the average disparity in metrics between the unlearned and retrained model, to compare their overall "closeness" [Huang et al., 2024], robustness to membership inference attacks (MIA), and computational efficiency through the number of iterations (Iters) required for unlearning. Ideally, the MIA score for any unlearning method should be close to 50% [Hu et al., 2022].

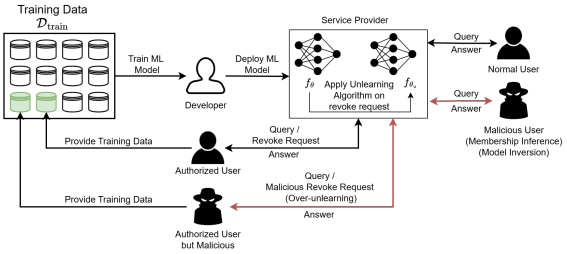


Figure 1: Overview of the MLaaS framework and potential black-box attacks that the unlearning algorithm must defend against to safeguard the model.

Assumptions: We assume the deployed model is well-trained, achieving high accuracy on $\mathcal{D}_{\text{train}}$ and reliable predictions on unseen data, consistent with the MLaaS setting. In this work, we focus on classification tasks. We also consider three distinct unlearning paradigms, sample-wise forgetting which entails unlearning a random subset of data points from $\mathcal{D}_{\text{train}}$, subclass-wise forgetting which requires unlearning all samples from $\mathcal{D}_{\text{train}}$ that belong to a particular subclass within a class and class-wise forgetting which unlearns all samples from $\mathcal{D}_{\text{train}}$ that belong to a single class.

MLaaS Framework: In the MLaaS paradigm, developers train proprietary models and deploy them on servers for commercialization. While the server handles model maintenance, including periodic updates, it lacks access to $\mathcal{D}_{\text{train}}$ and relies on $\mathcal{D}_{\text{test}}$ for monitoring model performance. To comply with data protection regulations, developers preselect an unlearning method executed by the server when authorized users submit data revocation requests for instances $x_i \in \mathcal{D}_u$. However, this opens avenues for malicious attacks. Black-box attacks, such as membership inference and model inversion, target $\mathcal{D}_{\text{train}}$ while authorized malicious users aim to degrade model utility through the exploitation of revocation rights by submitting corrupted inputs post-deployment. Fig. 1 illustrates the MLaaS framework and potential attack channels.

4 METHODOLOGY

Our proposed approach comprises two key components: *(i)* a probability redistribution module that redistributes the

predicted output $f_\theta(x_i)$ for each $x_i \in \mathcal{D}_u$ to simulate the removal of sample influence from f_θ and *(ii)* a regularized loss function that balances the performance across the retained dataset $\mathcal{D}_r$ and test data $\mathcal{D}_{\text{test}}$ while ensuring unlearning occurs effectively on $\mathcal{D}_u$.

4.1 ADAPTIVE PROBABILITY REDISTRIBUTION

We formulate unlearning for a given sample as reducing the probability mass assigned to the predicted class while redistributing it proportionally among the remaining classes. Let $p_i = \sigma\left(f_\theta(x_i)\right) \in \mathbb{R}^c$ where $\sigma(\cdot)$ is the Softmax function.

$$\begin{aligned} \widehat{p}_i^j &= \begin{cases} p_i^j(1-\alpha) & \text{if } j = \widehat{y}, \\ \frac{p_i^j}{\sum_{j,j\neq\widehat{y}_i} p_i^j}(1-\alpha)p_i^{\widehat{y}_i} + p_i^j & \text{otherwise} \end{cases} \\ \widehat{y}_i &= \underset{j\in\{1,\cdots,c\}}{\arg\max}\ p_i^j \end{aligned} \tag{1}$$

where $\widehat{y}_i$ is the predicted class for x_i while $\alpha \in [0,1]$ is a hyper-parameter controlling the extent of unlearning.

4.2 LOSS FUNCTION

Achieving effective unlearning requires fine-tuning θ to minimize the influence of samples in $\mathcal{D}_u$, while simultaneously preserving its performance on $\mathcal{D}_r$ and $\mathcal{D}_{\text{test}}$. To this end, we propose the following regularized loss function,

$$\text{KL}\left(\widehat{p}_i \parallel f_{\theta^t}(x_i)\right) + \sum_{l=1}^{L} \lambda_l \|\theta_l - \theta_l^t\|_F^2 \tag{2}$$

Here λ_l is a layer-wise regularization hyper-parameter while θ^t represents the parameters of the unlearned model at iteration t with $\theta^0 = \theta$. Each layer is assigned a unique λ to account for the heterogeneous importance of layers in large deep neural networks Zhang et al. [2022]. Section 6.3 outlines a straightforward method for determining λ values based on layer depth. The term $\text{KL}\left(\widehat{p}_i \parallel f_{\theta^t}(x_i)\right)$ ensures that f_θ is updated to align with the modified target distribution $\widehat{p}_i$, effectively reducing the influence of samples in $\mathcal{D}_u$. The term $\lambda_l\|\theta_l - \theta_l^t\|_F^2$ is a regularizer on the distance between the original pre-trained network θ and the updated parameters θ^t at iteration t in euclidean space which preserves model generalization on $\mathcal{D}_r$.

4.3 UNDERSTANDING AND OPTIMIZING HYPER-PARAMETER DYNAMICS

The balance between unlearning and retention is controlled by the hyper-parameters α and λ_l, respectively. Therefore, effective unlearning hinges on a precise understanding of hyper-parameter selection and their dynamic interplay.

Balancing Stability and Unlearning: The hyper-parameter α governs the extent of unlearning on $\mathcal{D}_u$. Initializing α close to 1 causes significant reductions in the predicted probabilities of the target class, resulting in a large deviation from the original predictions. This results in large losses and significant weight updates that push the model parameters far from their initial state, thereby degrading performance on $\mathcal{D}_r$. Conversely, initializing α close to 0 results in negligible changes to the predicted probabilities, rendering unlearning ineffective. To strike an effective balance, we propose using a cosine annealing Loshchilov and Hutter [2017] inspired dynamic update rule for α that gradually increases its value during training, enabling controlled unlearning while maintaining stability. Specifically, α is updated at each training iteration t as follows,

$$\alpha^t = \alpha_{\max} - \frac{1}{2}\left(\alpha_{\max} - \alpha_{\min}\right)\left(1 + \cos\left(\frac{t}{T-1}\pi\right)\right) \tag{3}$$

Here α^t represents the value of α at iteration t, $\alpha_{\min}$ and $\alpha_{\max}$ are hyper-parameters defining the the range of α and T denotes the total number of iterations. The proposed schedule begins with $\alpha^0 = \alpha_{\min}$ and gradually increases as training progresses until $\alpha^T = \alpha_{\max}$.

Preserving Generalization: The hyper-parameter λ_l controls the regularization that preserves model performance on $\mathcal{D}_r$. If initialized too high, λ_l overly constrains parameter updates, preventing effective unlearning of $\mathcal{D}_u$. Conversely, initializing it too low allows the parameters to drift excessively, leading to significant degradation in generalization. To ensure a smooth balance, we dynamically adjust λ_l during training. The update rule for λ_l at iteration t is,

$$\lambda_l^t = \lambda_{l_{\max}} - \frac{1}{2}\left(\lambda_{l_{\max}} - \lambda_{l_{\min}}\right)\left(1 + \cos\left(\frac{t}{T-1}\pi\right)\right) \tag{4}$$

Here, λ_l^t represents the value of λ_l at iteration t, $\lambda_{l_{\min}}$ and $\lambda_{l_{\max}}$ are hyperparameters defining the range of λ_l and T denotes the total number of iterations. Starting with $\lambda_l^0 = \lambda_{l_{\min}}$, this schedule allows flexibility in the early stages of training to prioritize unlearning. By the end, $\lambda_l^T = \lambda_{l_{\max}}$ reinforces constraints, pulling parameters closer to their original state and recovering performance on $\mathcal{D}_r$.

Interplay of α^t and λ_l^t: The dynamic interaction between α^t and λ_l^t is crucial for balancing unlearning and performance. Early in training, the gradual increase in α_t allows the network to incrementally update its weights. During this phase, λ_l^t exerts minimal influence, allowing the network to slowly unlearn $\mathcal{D}_u$. Even near the end of training, when α^t is close to $\alpha_{\max}$, the relative change in the class-wise target probabilities, in between iterations, is small, ensuring that parameter updates due to unlearning are never aggressive. On the contrary, the increase in λ_l^t forces the model weights to move closer to the original parameters thus allowing the network to recover performance on $\mathcal{D}_r$. This synergy between α^t and λ_l^t ensures that the model achieves effective unlearning without compromising its overall utility.

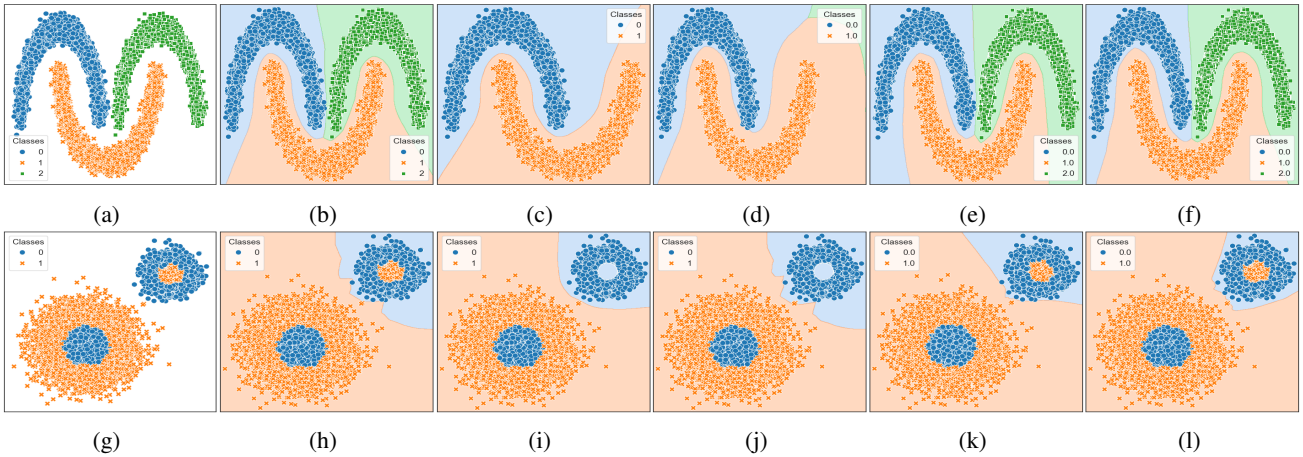


Figure 2: The non-linear synthetic datasets generated for the controlled simulations and their corresponding decision boundaries across multiple tasks. Fig. 2a, is the Moons dataset and Fig. 2g is the Blobs dataset. Figs. 2b and 2h are the visualization of corresponding decision boundaries. Figs. 2c and 2d are the retrained and unlearned models' decision boundaries, respectively, for class-wise forgetting on the Moons dataset. Figs. 2e and 2f are the retrained and unlearned models' decision boundaries, respectively, for sample-wise forgetting on the Moons dataset. Figs. 2i and 2j are the retrained and unlearned models' decision boundaries, respectively, for subclass-wise forgetting on the Blobs dataset. Figs. 2k and 2l are the retrained and unlearned models' decision boundaries, respectively, for sample-wise forgetting on the Blobs dataset.

5 CONTROLLED SIMULATIONS ON SYNTHETIC DATA

Existing metrics for machine unlearning offer valuable insights into post-forgetting performance but fail to fully capture how unlearning reshapes a model's decision-making behaviour. Since a model's classification performance fundamentally depends on its decision boundary, analyzing it directly reveals the impact of unlearning on generalization. Thus, we conduct controlled experiments on two synthetic datasets with known ground truths, applying SALSA under sample-wise, subclass-wise, and class-wise unlearning. This setup allows us to precisely evaluate how unlearning transforms the model's decision boundary, providing a deeper understanding of its generalization dynamics.

5.1 DATASET AND MODELS

We use two non-linear synthetic datasets, Moons scikit learn [2024b] and Blobs scikit learn [2024a], to evaluate SALSA across different unlearning paradigms. The Moons dataset comprises three classes, each with 5000 training and 500 test samples. A three-layer MLP, trained to convergence, learns the decision boundary shown in Fig. 2b. This dataset is used for class-wise and sample-wise unlearning. The Blobs dataset includes two classes, each with two distinct subclasses, reflecting hierarchical structures in real-world data. It contains 7500 training and 500 test samples per class. A three-layer MLP captures the decision boundary (Fig. 2h). This dataset is used for subclass-wise and sample-wise unlearning.

5.2 IMPLEMENTATION DETAILS

On the Moons dataset, class 3 is chosen for class-wise unlearning while the smaller sub-class of class 1 is chosen for subclass unlearning on the Blobs dataset. For sample-wise unlearning, we unlearn a random 10% subset of examples from both datasets. We report all results by averaging over three different runs for each experiment. Further details for reproducibility are provided in Section A of the Appendix.

Table 1: Simulation results on the Moons and Blobs datasets for sample-wise, subclass-wise and class-wise unlearning. Results have been averaged over three different runs.

Dataset	Task	Method	FA_{tr}	RA_{tr}	FA_{te}	RA_{te}	TA
Moons	Class wise	RT	00.0	100.0	00.0	100.0	-
		SALSA	00.0	98.8	00.6	98.4	-
	Sample wise	RT	100.0	100.0	-	-	100.0
		SALSA	100.0	100.0	-	-	100.0
Blobs	Subclass wise	RT	00.0	100.0	00.0	100.0	-
		SALSA	00.0	97.6	00.0	97.5	-
	Sample wise	RT	98.5	100.0	-	-	98.0
		SALSA	96.9	95.8	-	-	95.3

5.3 RESULTS

Across all scenarios, SALSA consistently matches retrained models in forgetting targeted information while preserving generalization (see Table 1 for fine-grained results).

Class-wise Unlearning: The ideal decision boundary

learned by the retrained model is visualized in Fig. 2c. SALSA closely replicates this boundary (Fig. 2d), effectively forgetting the class while maintaining overall generalization. The Average Discrepancy is 0.85%.

Subclass-wise Unlearning: The retrained boundary is visualized in Fig. 2i. SALSA reproduces this boundary (Fig. 2j) with a low Average Discrepancy of 1.23%.

Sample-wise Unlearning: Randomly removing 10% of training samples minimally affects the model's decision boundary. SALSA produces boundaries (Figs. 2f, 2l) nearly identical to those of the retrained models (Figs. 2e, 2k), with Average Discrepancies of 0% (Moons) and 2.8% (Blobs).

6 EXPERIMENTS ON STANDARD DATA

We assess the performance of our proposed approach on benchmark datasets, focusing on class-wise and sample-wise unlearning.

6.1 DATASET AND MODELS

Our experiments leverage three widely used image classification datasets with varying sizes, resolutions, and class distributions. **CIFAR10** [Krizhevsky et al., 2009] comprises $50,000$ training images and $10,000$ test images distributed uniformly over 10 classes. We use the ResNet18 model [He et al., 2016] for training on this dataset. The Street View House Numbers or **SVHN** [Netzer et al., 2011] is a real-world dataset with $73,257$ training images and $26,032$ test images across 10 classes. ResNet18 is employed for training on this dataset. The **TinyImageNet** [Yang, 2015] dataset consists of 200 classes, each comprising 500 training images and 50 test images, totaling $100,000$ training samples and $10,000$ test samples. Swin-T transformer [Liu et al., 2022] is chosen for training on this dataset.

6.2 BASELINES

We regard the retrained model (RT) as the oracle of approximate machine unlearning and compare SALSA against five state-of-the-art machine unlearning methods. These include SFTC [Perifanis et al., 2024], SalUn [Fan et al., 2024], FEMU [Tarun et al., 2024], LAF [Shen et al., 2024] and SFRon [Huang et al., 2024]. We also consider Fine-tuning (FT), a strong baseline, where f_θ is fine-tuned on $\mathcal{D}_r$. This is akin to catastrophic forgetting where fine-tuning without $\mathcal{D}_u$, may lead to unlearning.

6.3 IMPLEMENTATION DETAILS

According to Eqn. 4, determining $\lambda_{l_{\min}}$ and $\lambda_{l_{\max}}$ is essential to balance unlearning across network layers effectively. Recognizing that the initial layers of a neural network play a more critical role in learning [Zhang et al., 2022], we assign higher weights to these layers during unlearning which is achieved using an linear function.

$$\begin{aligned} \lambda_{l_{\min}} &= m \times (L - l + 1) + c \qquad \forall l \in \{1, \cdots, L\} \\ \lambda_{l_{\max}} &= \lambda_{l_{\min}} + \gamma \end{aligned} \tag{5}$$

where m and c are the scale and shift hyper-parameters that dictate how $\lambda_{l_{\min}}$ varies across the network depth. Meanwhile, γ determines how much λ_l^t increases as training progresses. We choose to forget the best-performing class in each dataset (classes $1, 4$ and 23 for CIFAR10, SVHN and TinyImageNet, respectively) as it typically causes the largest drop in performance, challenging all unlearning algorithms to match the performance of the oracle. Following [Shen et al., 2024], we forget 40% random subset of samples from the last 50% of classes for sample-wise unlearning in each dataset. On the CIFAR10 and SVHN datasets, the ResNet18 model is trained from scratch while we fine-tune an ImageNet [Russakovsky et al., 2015] pre-trained Swin-T transformer on TinyImageNet which closely follows MLaaS practice of fine-tuning strong foundation models on custom datasets. We report all results by averaging over three different runs for each experiment. Further implementation details for reproducibility are provided in Section A of the Appendix.

6.4 RESULTS

Under the metrics, $\text{FA}_{\text{tr}}, \text{RA}_{\text{tr}}, \text{FA}_{\text{te}}, \text{RA}_{\text{te}}$ and TA, the algorithm with the smallest discrepancy from the oracle (RT) is considered the best. Additionally, for the MIA and Iters metrics, the method achieving scores closest to 50% and 1, respectively, is considered optimal.

Class-wise and Sample-wise Unlearning: Table 2 showcases both class-wise and sample-wise unlearning performance across all datasets.

For **class-wise unlearning**, SALSA consistently outperforms all considered baselines across each dataset, achieving the lowest average discrepancy ($< 1.5\%$) from exact unlearning. This near-perfect approximation demonstrates that SALSA effectively emulates retraining without access to the original data, ensuring the preservation of model generalization. Moreover, SALSA achieves remarkable efficiency, being up to $25\times$ faster than the next fastest baseline, SFRon. Note that, due to LAF's high resource requirements, we were unable to evaluate it on TinyImageNet. In terms of privacy preservation, SALSA's MIA scores are consistently close to 50%, aligning with the gold standard for privacy. For instance, on SVHN and TinyImageNet, even RT exhibits slight membership leakage, with MIA scores deviating from 50%. In contrast, SALSA's near-ideal MIA scores showcase robust privacy preservation while maintaining high utility.

Table 2: Combined class-wise and sample-wise unlearning performance comparison. For class-wise unlearning, the class with the highest training accuracy is unlearned. For sample-wise unlearning, a random subset of 10% examples is unlearned. Values closest to RT under each metric are bolded, with the second best underlined. For MIA, scores nearest to 50% are bolded, with the second nearest underlined. For Iters, the method with the fewest iterations is bolded, with the second best underlined. Results are averaged over three runs.

		Class-wise Unlearning							Sample-wise Unlearning					
Dataset	Methods	FA_{tr}	RA_{tr}	FA_{te}	RA_{te}	Avg. D	MIA	Iters	FA_{tr}	RA_{tr}	TA	Avg. D	MIA	Iters
CIFAR10	RT	00.0	100	00.0	95.6	–	46.6	70K	93.3	100	92.6	–	60.4	56K
	FT	**00.0**	92.7	**00.0**	88.5	03.6	60.6	14K	99.8	99.9	95.0	3.0	64.2	11K
	SFTC	**00.0**	92.7	**00.0**	88.6	03.6	59.0	16K	**100**	**100**	**95.4**	3.2	66.4	15K
	SalUn	**00.0**	85.5	**00.0**	82.8	06.8	63.1	4K	**100**	**100**	95.3	3.1	66.2	3K
	LAF	**00.0**	60.2	**00.0**	**93.3**	10.5	62.4	5K	94.1	99.9	93.9	0.7	55.6	109K
	SFRon	**00.0**	74.5	**00.0**	73.2	12.0	62.0	2K	**100**	99.9	95.2	3.1	65.5	4K
	SALSA	00.3	**98.7**	00.5	92.0	**01.4**	**50.2**	**0.1K**	96.9	97.9	**95.4**	**2.9**	**51.1**	**0.02K**
SVHN	RT	00.0	99.4	00.0	96.7	–	62.6	70K	92.9	99.9	95.8	–	54.8	664K
	FT	**00.0**	96.0	**00.0**	95.6	01.1	56.7	374K	97.8	99.4	**96.4**	**2.0**	55.3	8K
	SFTC	**00.0**	95.7	**00.0**	95.4	01.3	59.5	20K	**99.8**	**99.8**	**96.4**	2.6	55.6	33K
	SalUn	**00.0**	93.3	**00.0**	93.0	02.4	63.2	22K	99.7	99.7	**96.4**	2.5	55.7	33K
	LAF	04.1	65.2	01.3	**96.4**	10.0	64.5	595K	00.0	98.7	60.2	43.2	33.5	595K
	SFRon	**00.0**	90.7	**00.0**	90.4	03.8	**53.1**	2K	97.8	99.1	96.2	**2.0**	55.7	2K
	SALSA	00.7	**99.6**	00.7	95.3	**00.7**	45.7	**0.2K**	97.1	98.1	91.2	3.5	**51.1**	**0.03K**
Tiny Image Net	RT	00.0	92.0	00.0	85.3	–	63.4	78K	85.81	97.8	85.81	–	58.06	78K
	FT	82.8	**92.3**	84.0	84.4	42.0	60.0	4K	67.8	77.6	69.3	18.0	55.3	3K
	SFTC	98.2	91.4	84.0	83.4	49.8	63.2	4K	76.6	77.6	70.1	14.8	56.9	4K
	SalUn	97.4	91.5	98.0	84.0	49.3	62.3	4K	75.2	76.8	70.4	15.4	56.9	4K
	LAF	–	–	–	–	–	–	–	–	–	–	–	–	–
	SFRon	**00.0**	91.4	**00.0**	84.7	00.3	39.3	1K	48.3	65.7	59.3	31.8	55.8	1K
	SALSA	00.2	91.5	**00.0**	**85.3**	**00.2**	**47.8**	**0.02K**	**89.7**	**91.2**	**84.9**	**3.6**	**49.1**	**0.08K**

For **sample-wise unlearning**, SALSA outperforms all considered baselines on CIFAR10 and TinyImageNet, achieving the lowest average discrepancy ($< 4\%$) from exact unlearning. On SVHN, SFRon achieves the lowest average discrepancy. However, our approach is $84\times$ faster, on average, than SFRon, the fastest considered baseline. Moreover, SALSA's MIA scores are near 50% on all datasets, aligning with the gold standard for privacy. The combined results demonstrate SALSA's superior efficiency across both unlearning scenarios while maintaining competitive performance metrics and robust privacy preservation.

7 MODEL INVERSION ATTACK

7.1 SETUP

To assess the robustness of SALSA, we evaluate its effectiveness against MIRROR [Tao et al., 2022], a state-of-the-art model inversion attack. We adopt the experimental setup of MIRROR, where a StyleGAN [Karras et al., 2019] pre-trained on the CelebA dataset [Liu et al., 2015] serves as the generator for the attack. An InceptionResNet-v1 [Szegedy et al., 2017] pre-trained on the VGGFace2 dataset [Cao et al., 2018] is targeted in the attack. The StyleGAN iteratively optimizes its generated images using a genetic algorithm [Bhandari et al., 1996], aiming to infer private data from the model in a black-box setting. The hyperparameters for unlearning are detailed in Section A of the Appendix.

7.2 RESULTS

To demonstrate the efficacy of SALSA against state-of-the-art black box model inversion attack, we select two visually similar classes from the VGGFace2 dataset (Figs. 3a and 3b). Without unlearning, MIRROR reconstructs facial and hair features of private training samples with striking fidelity, as shown in Figs. 3c and 3d. However, after unlearning with SALSA, MIRROR struggles to recover even rudimentary information about the forgotten classes, as demonstrated in Fig. 3e. A critical goal of unlearning is to ensure that data from retained classes remains unaffected. Fig. 3f confirms this, as the images generated by MIRROR for the retained class are nearly indistinguishable from the originals. Our re-

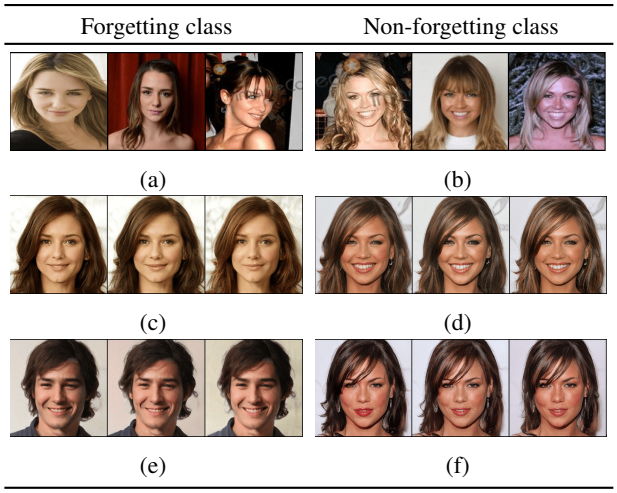


Figure 3: The black box model inversion attack variant of MIRROR. Figs. 3a and 3b are images taken from two visually similar classes in the VGGFace2 dataset and represent the ground truth. Figs. 3c and 3d are randomly selected batches of output generated by the inversion attack on the pre-trained network for each class before unlearning. Figs. 3e and 3f are randomly selected batches of output generated by the same attack model on the unlearned network.

sults highlight SALSA's robustness against model inversion attacks, successfully erasing private data while maintaining the integrity of retained classes.

8 MALICIOUS UNLEARNING

8.1 SETUP

Malicious Unlearning (Over-unlearning) simulates a black-box attack in MLaaS, where the server has no knowledge of the unlearning request. We specifically implement subset over-unlearning II as proposed by Hu et al. [2024a]. For each dataset, we shift 50% of the best-performing class' samples just across the model's decision boundary, pushing them toward the second-highest-performing class. These adversarially modified samples form the unlearning request.

8.2 RESULTS

Hu et al. [2024a] showed that over-unlearning can significantly degrade test accuracy (TA). However, SALSA remains remarkably robust, experiencing nearly no drop in performance on CIFAR10 and TinyImageNet (Table 3). On SVHN, however, TA drops by over 10%. We attribute this to the high visual similarity between the top two performing classes, digits 1 and 4, making decision boundaries more susceptible to perturbations. These results highlight SALSA's resilience against adversarial unlearning in most MLaaS settings.

Table 3: Test accuracy (TA) for Malicious Unlearning (over-unlearning) in contrast to normal (benign) unlearning.

Dataset	Benign Unlearn	Malicious Unlearn
CIFAR10	88.57	88.42
SVHN	86.71	74.06
TinyImageNet	86.64	86.71

9 ANALYSIS STUDY

We investigate different aspects of SALSA through the lens of class-wise unlearning.

Visualizing the Unlearning: We leverage GradCAM [Selvaraju et al., 2017] to visualize how the Swin-T transformer attends to images from both unlearned and retained classes in TinyImageNet. Fig. 4 shows activation maps before and after applying SALSA, highlighting the forget set (class 23) and three random retained samples. Post-unlearning, the model no longer focuses on key regions, indicating the effective removal of class-specific information. This visualization provides intuitive evidence of SALSA's ability to successfully unlearn without compromising generalization.

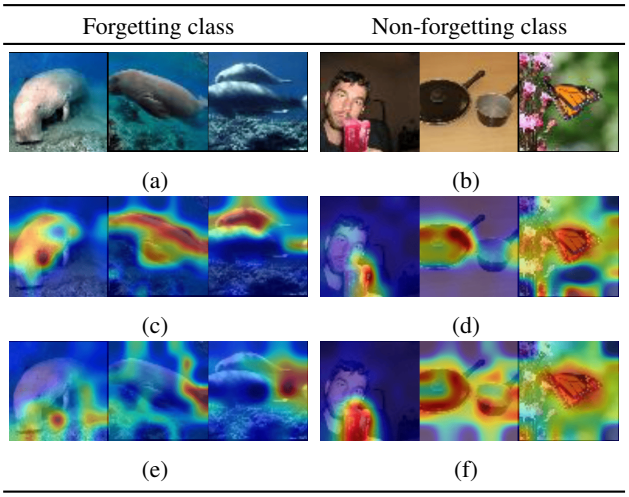
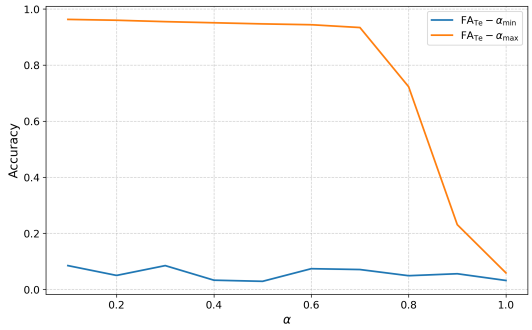
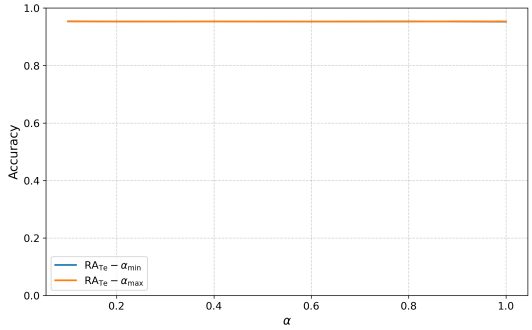


Figure 4: Swin-T transformer activation maps on TinyImageNet. Fig. 4a shows a random batch from the forget set, while Fig. 4b shows random images from the retained set. Figs. 4c and 4d depict activation maps before unlearning, whereas Figs. 4e and 4f demonstrate the same map but after unlearning.

Effect of varying $\alpha_{\min}$ and $\alpha_{\max}$: Fixing $\alpha_{\max} = 1$, we expect FA_{te} to start low and approach 0 as $\alpha_{\min}$ increases from 0 to 1. This is because smaller $\alpha_{\min}$ values delay most of the probability mass redistribution until the final stages of unlearning. Similarly, fixing $\alpha_{\min} = 0$ and varying $\alpha_{\max}$ in $[0, 1]$ should amplify this effect, as class-wise unlearning demands $\alpha_{\max} \to 1$. Fig. 5 confirms this trend for ResNet18 on CIFAR10. Notably, SALSA remains robust, i.e., RA_{te}



(a) Change in forgetting accuracy on the test set, (FA_{te})



(b) Change in retained accuracy on the test set, (RA_{te})

Figure 5: Accuracy trends for retain and forget classes on test sets with varying alpha values.

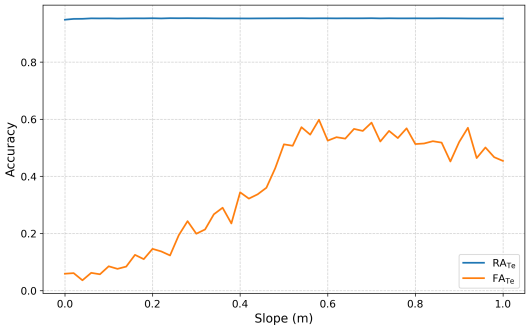


Figure 6: Accuracy variation with the slope (m), showing its effect on retention and forgetting.

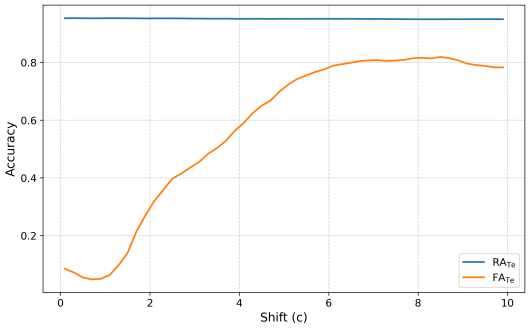


Figure 7: Accuracy dependence on the shift (c), illustrating how offset adjustments influence performance.

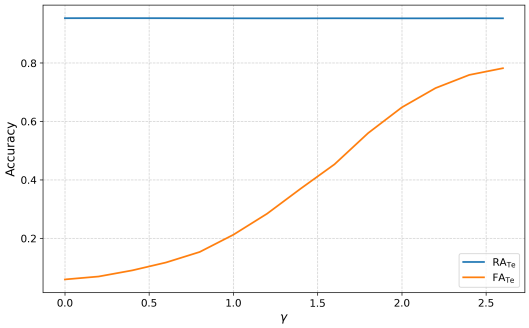


Figure 8: Accuracy trends with respect to γ, highlighting its role in modulating class retention and forgetting.

remains stable while FA_{te} steadily declines.

Effect of varying m, c, γ: Increasing m amplifies weight penalization near the input layers, while higher c enforces stronger regularization across all layers. Larger γ further intensifies weight penalization as unlearning progresses. In all cases, we expect the amount of unlearning to reduce, i.e., FA_{te} should rise due to increased regularization. Figs. 6, 7, and 8 validate this trend for ResNet18 on CIFAR10. Once again, SALSA remains robust to changes in RA_{te} with respect to the variation in the regularization hyperparameters.

10 CONCLUSION

In this work, we present SALSA, a Scalable, Adaptive and Label-Agnostic Scalable Algorithm for machine unlearning tailored for classification tasks in the MLaaS scenario. SALSA redistributes model output probabilities for samples that need to be forgotten. Thereafter, it employs a novel loss function that minimizes the divergence between predicted and redistributed probabilities while maintaining minimum distance from model initialization. This ensures simultaneous unlearning and generalization. Our approach is label independent and requires only the samples to be forgotten, for unlearning which makes SALSA exceptionally fast, achieving at least $25\times$ and $84\times$ faster class-wise and sample-wise unlearning, respectively, than the fastest considered baseline. Extensive experiments on benchmark and synthetic datasets show that SALSA achieves the closest approximation to exact unlearning. Through rigorous evaluations against state-of-the-art black box attacks, we demonstrate SALSA's resilience to privacy and security threats. By balancing utility and privacy at scale, SALSA marks a significant step forward in practical, privacy-preserving unlearning for MLaaS and sensitive data management.

References

Shengwei An, Guanhong Tao, Qiuling Xu, Yingqi Liu, Guangyu Shen, Yuan Yao, Jingwei Xu, and Xiangyu Zhang. Mirror: Model inversion for deep learning network with high fidelity. In *Proceedings of the 29th Network and Distributed System Security Symposium*, 2022.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Dinabandhu Bhandari, CA Murthy, and Sankar K Pal. Genetic algorithm with elitist model and its convergence. *International journal of pattern recognition and artificial intelligence*, 10(06):731–747, 1996.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *42nd IEEE Symposium on Security and Privacy, SP 2021, San Francisco, CA, USA, 24-27 May 2021*, pages 141–159. IEEE, 2021. doi: 10.1109/SP40001.2021.00019.

Jonathan Brophy and Daniel Lowd. Machine unlearning for random forests. In *International Conference on Machine Learning*, pages 1092–1104. PMLR, 2021.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020a.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020b.

Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *13th IEEE International Conference on Automatic Face & Gesture Recognition, FG 2018, Xi'an, China, May 15-19, 2018*, pages 67–74. IEEE Computer Society, 2018. doi: 10.1109/FG.2018.00020.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pages 463–480. IEEE, 2015.

Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 5253–5270, 2023.

Min Chen, Zhikun Zhang, Tianhao Wang, Michael Backes, Mathias Humbert, and Yang Zhang. When machine unlearning jeopardizes privacy. In *Proceedings of the 2021 ACM SIGSAC conference on computer and communications security*, pages 896–911, 2021.

Ruizhe Chen, Jianfei Yang, Huimin Xiong, Jianhong Bai, Tianxiang Hu, Jin Hao, Yang Feng, Joey Tianyi Zhou, Jian Wu, and Zuozhu Liu. Fast model debias with machine unlearning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023.

Zehua Ding, Youliang Tian, Guorong Wang, Jinbo Xiong, Jinchuan Tang, and Jianfeng Ma. Membership inference attacks via spatial projection-based relative information loss in mlaas. *Information Processing & Management*, 62(1):103947, 2025. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2024.103947.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Min Du, Zhi Chen, Chang Liu, Rajvardhan Oak, and Dawn Song. Lifelong anomaly detection through unlearning. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, CCS 2019, London, UK, November 11-15, 2019*, pages 1283–1297. ACM, 2019. doi: 10.1145/3319535.3363226.

Chongyu Fan, Jiancheng Liu, Yihua Zhang, Eric Wong, Dennis Wei, and Sijia Liu. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1322–1333, 2015.

Matthew Fredrikson, Eric Lantz, Somesh Jha, Simon Lin, David Page, and Thomas Ristenpart. Privacy in pharmacogenetics: An {End-to-End} case study of personalized warfarin dosing. In *23rd USENIX security symposium (USENIX Security 14)*, pages 17–32, 2014.

Ji Gao, Sanjam Garg, Mohammad Mahmoody, and Prashant Nalini Vasudevan. Deletion inference, reconstruction, and compliance in machine (un) learning. *Proceedings on Privacy Enhancing Technologies*, 2022.

Laura Graves, Vineel Nagisetty, and Vijay Ganesh. Amnesiac machine learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11516–11524, 2021.

Chuan Guo, Tom Goldstein, Awni Y. Hannun, and Laurens van der Maaten. Certified data removal from machine learning models. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 3832–3842. PMLR, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.90.

Hongsheng Hu, Zoran Salcic, Lichao Sun, Gillian Dobbie, Philip S Yu, and Xuyun Zhang. Membership inference attacks on machine learning: A survey. *ACM Computing Surveys (CSUR)*, 54(11s):1–37, 2022.

Hongsheng Hu, Shuo Wang, Jiamin Chang, Haonan Zhong, Ruoxi Sun, Shuang Hao, Haojin Zhu, and Minhui Xue. A duty to forget, a right to be assured? exposing vulnerabilities in machine unlearning services. In *31st Annual Network and Distributed System Security Symposium, NDSS 2024, San Diego, California, USA, February 26 - March 1, 2024*. The Internet Society, 2024a.

Hongsheng Hu, Shuo Wang, Tian Dong, and Minhui Xue. Learn what you want to unlearn: Unlearning inversion attacks against machine unlearning. In *IEEE Symposium on Security and Privacy, SP 2024, San Francisco, CA, USA, May 19-23, 2024*, pages 3257–3275. IEEE, 2024b. doi: 10.1109/SP54263.2024.00248.

Yuke Hu, Jian Lou, Jiaqi Liu, Wangze Ni, Feng Lin, Zhan Qin, and Kui Ren. Eraser: Machine unlearning in mlaas via an inference serving-aware approach. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pages 3883–3897, 2024c.

Zhehao Huang, Xinwen Cheng, JingHao Zheng, Haoran Wang, Zhengbao He, Tao Li, and Xiaolin Huang. Unified gradient-based machine unlearning with remain geometry enhancement. *CoRR*, abs/2409.19732, 2024. doi: 10.48550/ARXIV.2409.19732.

Zachary Izzo, Mary Anne Smart, Kamalika Chaudhuri, and James Zou. Approximate data deletion from machine learning models. In *International Conference on Artificial Intelligence and Statistics*, pages 2008–2016. PMLR, 2021.

Mostafa Kahla, Si Chen, Hoang Anh Just, and Ruoxi Jia. Label-only model inversion attacks via boundary repulsion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15045–15053, 2022.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4401–4410. Computer Vision Foundation / IEEE, 2019. doi: 10.1109/CVPR.2019.00453.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer V2: scaling up capacity and resolution. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 11999–12009. IEEE, 2022. doi: 10.1109/CVPR52688.2022.01170.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 3730–3738. IEEE Computer Society, 2015. doi: 10.1109/ICCV.2015.425.

Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with warm restarts. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

Zhaobo Lu, Hai Liang, Minghao Zhao, Qingzhe Lv, Tiancai Liang, and Yilei Wang. Label-only membership inference attacks on machine unlearning without dependence of posteriors. *International Journal of Intelligent Systems*, 37(11):9424–9441, 2022.

Adithyan M Nair, Akshit Sudheer Kumar, Devakrishna Sanil Kumar, and Anjali T. Selective unlearning in face recognition: Forgetting faces without compromising accuracy. In *2023 3rd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, pages 211–216, 2023. doi: 10.1109/ICIMIA60377.2023.10426386.

Deepak Narayanan, Mohammad Shoeybi, Jared Casper, Patrick LeGresley, Mostofa Patwary, Vijay Korthikanti, Dmitri Vainbrand, Prethvi Kashinkunti, Julie Bernauer, Bryan Catanzaro, et al. Efficient large-scale language model training on gpu clusters using megatron-lm. In *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 1–15, 2021.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Baolin Wu, Andrew Y Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, page 4. Granada, 2011.

OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. doi: 10.48550/ARXIV.2303.08774.

Vasileios Perifanis, Efstathios Karypidis, Nikos Komodakis, and Pavlos Efraimidis. Sftc: Machine unlearning via selective fine-tuning and targeted confusion. In *Proceedings of the 2024 European Interdisciplinary Cybersecurity Conference*, EICC '24, page 29–36, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400716515. doi: 10.1145/3655693.3655697.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR, 2023.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

scikit learn. make_gaussian_quantiles function to generate the blobs dataset. `https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_gaussian_quantiles.html`, 2024a.

scikit learn. make_moons function to generate the moons dataset. `https://scikit-learn.org/dev/modules/generated/sklearn.datasets.make_moons.html`, 2024b.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

Shaofei Shen, Chenhao Zhang, Yawen Zhao, Alina Bialkowski, Weitong Chen, and Miao Xu. Label-agnostic forgetting: A supervision-free unlearning in deep models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.

Galit Shmueli, Peter C Bruce, Kuber R Deokar, and Nitin R Patel. *Machine learning for business analytics: Concepts, techniques, and applications with analytic solver data mining*. John Wiley & Sons, 2023.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18, 2017a. doi: 10.1109/SP.2017.41.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017b.

Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

Guanhong Tao, Qiuling Xu, Yingqi Liu, Guangyu Shen, Shengwei An, Jingwei Xu, Xiangyu Zhang, and Yuan Yao. MIRROR: model inversion for deep learningnetwork with high fidelity. In *29th Annual Network and Distributed System Security Symposium, NDSS 2022, San Diego, California, USA, April 24-28, 2022*. The Internet Society, 2022.

Ayush K. Tarun, Vikram S. Chundawat, Murari Mandal, and Mohan Kankanhalli. Fast yet effective machine unlearning. *IEEE Transactions on Neural Networks and Learning Systems*, 35(9):13046–13055, 2024. doi: 10.1109/TNNLS.2023.3266233.

Anvith Thudi, Gabriel Deza, Varun Chandrasekaran, and Nicolas Papernot. Unrolling SGD: understanding factors influencing machine unlearning. In *7th IEEE European Symposium on Security and Privacy, EuroS&P 2022, Genoa, Italy, June 6-10, 2022*, pages 303–319. IEEE, 2022a. doi: 10.1109/EUROSP53844.2022.00027.

Anvith Thudi, Gabriel Deza, Varun Chandrasekaran, and Nicolas Papernot. Unrolling sgd: Understanding factors

influencing machine unlearning. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*, pages 303–319. IEEE, 2022b.

Cheng-Long Wang, Mengdi Huai, and Di Wang. Inductive graph unlearning. In *32nd USENIX Security Symposium, USENIX Security 2023, Anaheim, CA, USA, August 9-11, 2023*, pages 3205–3222. USENIX Association, 2023.

Alexander Warnecke, Lukas Pirch, Christian Wressnegger, and Konrad Rieck. Machine unlearning of features and labels. In *30th Annual Network and Distributed System Security Symposium, NDSS 2023, San Diego, California, USA, February 27 - March 3, 2023*. The Internet Society, 2023.

Yuhuai Wu, Yuri Burda, Ruslan Salakhutdinov, and Roger Grosse. On the quantitative analysis of decoder-based generative models. In *International Conference on Learning Representations*, 2022.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. The rise and potential of large language model based agents: A survey. *arXiv preprint arXiv:2309.07864*, 2023.

Jie Xu, Zihan Wu, Cong Wang, and Xiaohua Jia. Machine unlearning: Solutions and challenges. *IEEE Trans. Emerg. Top. Comput. Intell.*, 8(3):2150–2168, 2024. doi: 10.1109/TETCI.2024.3379240.

Haonan Yan, Xiaoguang Li, Ziyao Guo, Hui Li, Fenghua Li, and Xiaodong Lin. ARCANE: an efficient architecture for exact machine unlearning. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 4006–4013. ijcai.org, 2022. doi: 10.24963/IJCAI.2022/556.

Xuan Yang. Tiny imagenet visual recognition challenge. 2015.

Ziqi Yang, Jiyi Zhang, Ee-Chien Chang, and Zhenkai Liang. Neural network inversion in adversarial setting via background knowledge alignment. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 225–240, 2019.

Chiyuan Zhang, Samy Bengio, and Yoram Singer. Are all layers created equal? *J. Mach. Learn. Res.*, 23:67:1–67:28, 2022.

Juexiao Zhou, Haoyang Li, Xingyu Liao, Bin Zhang, Wenjia He, Zhongxiao Li, Longxi Zhou, and Xin Gao. A unified method to revoke the private data of patients in intelligent healthcare with audit to forget. *Nature Communications*, 14(1):6255, 2023.

SALSA: A Secure, Adaptive and Label-Agnostic Scalable Algorithm for Machine Unlearning (Supplementary Material)

Owais Makroo[*1] **Atif Hassan**[*2] **Swanand Khare**[1]

[1]Department of Mathematics, IIT Kharagpur, Kharagpur, West Bengal, India
[2]Department of Artificial Intelligence, IIT Kharagpur, Kharagpur, West Bengal, India

A CHOICE OF HYPER-PARAMETERS

A.1 UNLEARNING

Table 4: Hyperparameters for different datasets and model combinations for both class-wise and sample-wise unlearning.

Dataset	Model	Task	Epochs	Batch Size	lr	$\alpha_{\min}$	$\alpha_{\max}$	slope (m)	shift (c)	γ
CIFAR10	ResNet18	Class	10	512	2.0e-4	0.1	0.9	0.1	1	0.4
		Sample	1	512	1.0e-5	0.1	0.3	0.01	0.1	0.2
SVHN	ResNet18	Class	10	512	2.0e-4	0.1	0.9	0.1	1	0.4
		Sample	1	512	1.0e-5	0.1	0.3	0.01	0.1	0.2
TinyImageNet	SwinT	Class	10	32	1.0e-3	1-1e-6	1	4.0e-3	1.0e-2	3.0e-2
		Sample	1	32	9.0e-4	0.3	0.6	5.0e-4	5.0e-3	1.0e-2
Blobs	MLP	Subset	50	64	2.0e-4	0.3	1	0.09	0.05	0.04
		Sample	20	64	2.0e-6	0.01	0.05	0.01	0.05	0.01
Moons	MLP	Class	20	64	2.0e-4	0.3	1	0.07	0.01	0
		Sample	20	64	2.0e-6	0.01	0.05	0.01	0.05	0.01
VGGFACE2	Inception ResnetV1	Class	10	16	1.0e-3	1-1e-8	1	1.0e-5	1.0e-4	1.0e-2

A.2 ORIGINAL TRAINING

Table 5: Hyperparameters for different dataset and model combinations.

Dataset	Model	Epochs	Batch Size	lr
CIFAR10	ResNet18	200	128	0.1
SVHN	ResNet18	200	128	0.1
TinyImageNet	SwinT	200	128	0.1
Blobs	MLP	75	64	0.01
Moons	MLP	50	64	0.01