Rethinking Adversarial Robustness in the Context of the Right to be Forgotten

Chenxu Zhao [*1] **Wei Qian** [*1] **Yangyi Li** [1] **Aobo Chen** [1] **Mengdi Huai** [1]

Abstract

The past few years have seen an intense research interest in the practical needs of the "right to be forgotten", which has motivated researchers to develop *machine unlearning* methods to unlearn a fraction of training data and its lineage. While existing machine unlearning methods prioritize the protection of individuals' private data, they overlook investigating the unlearned models' susceptibility to adversarial attacks and security breaches. In this work, we uncover a novel security vulnerability of machine unlearning based on the insight that adversarial vulnerabilities can be bolstered, especially for adversarially robust models. To exploit this observed vulnerability, we propose a novel attack called ***Adv**ersarial **U**nlearning **A**ttack* (AdvUA), which aims to generate a small fraction of malicious unlearning requests during the unlearning process. AdvUA causes a significant reduction of adversarial robustness in the unlearned model compared to the original model, providing an entirely new capability for adversaries that is infeasible in conventional machine learning pipelines. Notably, we also show that AdvUA can effectively enhance model stealing attacks by extracting additional decision boundary information, further emphasizing the breadth and significance of our research. We also conduct both theoretical analysis and computational complexity of AdvUA. Extensive numerical studies are performed to demonstrate the effectiveness and efficiency of the proposed attack.

1. Introduction

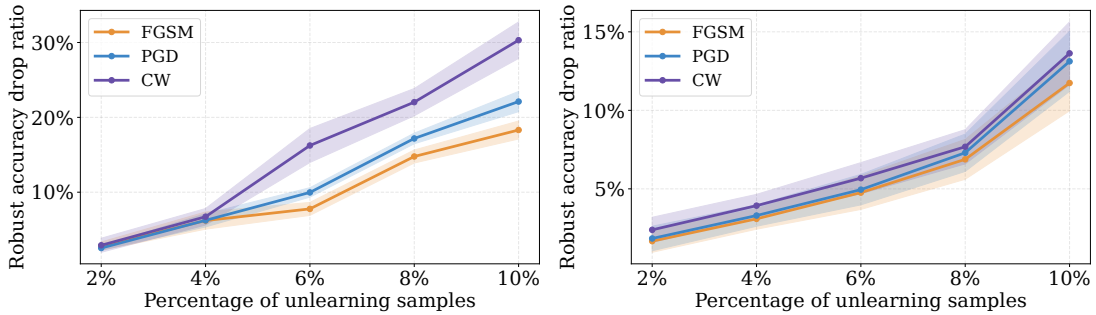
In recent years, many countries have raised concerns about protecting personal privacy. In practice, users may choose to have their data completely removed from a system, especially sensitive systems such as those do with finance or healthcare (Nguyen et al., 2022). Recent regulations (e.g., the well-known European Union's GDPR (Voigt & Von dem Bussche, 2017)) now compel organizations to give users "the right to be forgotten", i.e., the right to have all or part of their data deleted from a system on request (Liu et al., 2024b; Che et al., 2023; Chourasia & Shah, 2023; Zhang et al., 2022b; Graves et al., 2021; Chen et al., 2021; Wang et al., 2023; Brophy & Lowd, 2021; Liu et al., 2023a). The most straightforward approach is to retrain the model on all data except the portion that has been removed, but this approach is in general impractical since the computational resources consumed are usually costly. Thus, aiming to efficiently remove data as well as their contribution to the model, a new machine learning privacy protection research direction has emerged, called *machine unlearning*.

Numerous research efforts have been dedicated to addressing the challenge of data removal in inefficient retraining. Two prominent research fields that have emerged in this context are *exact unlearning* (Yan et al., 2022; Yu et al., 2022b; Brophy & Lowd, 2021; Bourtoule et al., 2021b; Qian et al., 2022) and *approximate unlearning* (Tarun et al., 2023; Jia et al., 2023; Chen et al., 2022; Gupta et al., 2021; Neel et al., 2021; Sekhari et al., 2021). Exact unlearning aims to completely reverse the effects of the previously learned data points. Instead of aiming for a perfect reversal of the learned data, approximate unlearning seeks to achieve a reasonably close approximation. This can be beneficial in situations where complete unlearning is computationally expensive or impractical, providing a compromise between removal efficiency and performance retention.

However, current machine unlearning methods exhibit an important limitation as they primarily concentrate on efficiently removing the effects of specific data instances from a trained machine learning model. Consequently, it remains uncertain whether these existing techniques might unexpectedly influence the adversarial robustness of the associated machine learning models. Note that adversarial robustness, in the context of machine learning, refers to the ability of a trained model to maintain its accuracy and performance even when it is exposed to deliberately perturbed input data known as adversarial examples (Li et al., 2019; Goodfellow et al., 2014; Szegedy et al., 2013; Zhu et al., 2021). Despite

[*]Equal contribution [1]Department of Computer Science, Iowa State University, United States. Correspondence to: Mengdi Huai <mdhuai@iastate.edu>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.



(a) SVHN (b) CIFAR-10

Figure 1. Robustness degradation of defended robust models against FGSM, PGD, and CW attacks after randomly removing different percentages of training samples on SVHN and CIFAR-10.

the great importance of studying adversarial robustness from the selective forgetting perspective, there is no existing work exploring the adversarial robustness properties of the unlearned models in the context of the right to be forgotten. Therefore, it is natural to ask

Q1: Does unlearning amplify the vulnerability of unlearned models to adversarial attacks?

To answer this question, we first conducted initial experiments to investigate adversarial robustness degradation of the defended and undefended models[1] by randomly deleting some training samples in the context of machine unlearning. In the experiments, the defended robust models are constructed using adversarial training (Madry et al., 2018; Goodfellow et al., 2014), which proves to be the most effective method against adversarial attacks (Bai et al., 2021; Pang et al., 2020; Maini et al., 2020). The randomly chosen training samples are removed using the first-order based unlearning method (Warnecke et al., 2023). The experimental results are presented in Figure 1, with additional details available in the full version of this paper. Notably, these experimental results show that the random removal strategy substantially amplifies the adversarial vulnerabilities of defended robust models. Additionally, in our experiments, we also found that compared with naturally undefended models, adversarially robust models are indeed *more susceptible* to malicious unlearning samples. The reason is that existing robust training methods heavily rely on the training data to enhance the model's robustness against adversarial attacks. Motivated by this, it is important to further investigate that

Q2: Are there specific requested unlearning samples that play a pivotal role in generating these new successful adversarial examples, which were unattainable before the unlearning process?

[1]Throughout the paper, we use "undefended (natural) model" and "defended (robust) model" to denote the machine learning model with natural training algorithm and robust training algorithm, respectively.

This work aims to provide answers to the above questions and highlight a potential adversarial attack vulnerability in the unlearning process – an adversary can make use of the unlearning pipeline to craft malicious unlearning requests to achieve his desired adversarial attack goals. Note that like traditional data poisoning attacks, the recent works (Hu et al., 2023; Qian et al., 2023; Di et al., 2022) still focus on how to poison the training data, and fail to study the impact of unlearning on the models' vulnerabilities to adversarial attacks. Therefore, in this paper, we aim to conduct an investigation into the adversarial risks associated with exercising the "right to be forgotten" from machine learning models, without altering the training data.

Our Contributions. To develop an understanding of adversarial risks associated with the process of unlearning, we in this paper present a novel attack named ***Adversarial Unlearning Attack*** (***AdvUA***), which exploits the unlearning pipeline to increase adversarial vulnerabilities. The key idea behind AdvUA is to select unlearning samples that are not only in close proximity to the target victim samples but also align with the adversarial attack directions. Specifically, in our proposed method, we first design a distance evaluation metric to estimate the space-filling capability of the region surrounding the target victim samples in the representation space. Then, based on the insight that not all nearest neighbor samples are equally critical for performing adversarial unlearning attacks, we propose a direction alignment loss to closely match the adversarial attack with the unlearning attack. *Our method is orthogonal to existing approaches on adversarial attacks and can be easily integrated with them to create advanced adversarial attack strategies.* Furthermore, it is worth highlighting that AdvUA can also *bolster the effectiveness of model stealing attacks* by extracting more decision boundary information, underlining the extensive scope and importance of our research.

We also empirically demonstrate the attack effectiveness and computational efficiency of AdvUA on various bench-

marks, including CIFAR-10 (Krizhevsky & Hinton, 2010) and ImageNet (Deng et al., 2009), against various robust learning methods in Section 4. Our evaluation also indicates that AdvUA performs well across different model architectures and machine unlearning methods. Overall, by conducting this study, we aim to shed light on the potential consequences of applying machine unlearning techniques to adversarially robust models and to gain insights into the interplay between data removal and model robustness against adversarial attacks. Ultimately, our findings will contribute to a more comprehensive understanding of the implications of machine unlearning in the context of adversarial machine learning and its implications for real-world applications.

2. Background and Related Work

Notations and Machine Unlearning. Let $S = \{(x_i, y_i)\}_{i=1}^n$ denote the dataset, where $x_i \in \mathcal{X} \subset \mathbb{R}^d$ is a d-dimensional feature and $y_i \in \mathcal{Y} = \{1, \cdots, C\}$. Let us suppose that a C-label classifier $F(W) : \mathbb{R}^d \to \mathbb{R}^C$ labels a sample x as $\arg\max_{c \in \mathcal{Y}} F(x; W)[c]$, where $W \in \mathcal{W}$ represents the parameters of F. For F, we denote H as the representation learning function and G as the final prediction head, i.e., $F(W) = G \circ H$. Given the learning algorithm L and S, the model owner can train a model $F(W^*)$ such that $F(W^*)$ achieves a low empirical loss. In machine unlearning, users can submit data removal requests $S_u \subset S$ to eliminate the influence of S_u from W^*, leading to the creation of the unlearned model $W^u \in \mathcal{W}$. Note that machine unlearning can be divided into: *exact* and *approximate*. Below, we outline the definitions of existing approximate and exact unlearning techniques (Warnecke et al., 2023; Gupta et al., 2021; Neel et al., 2021; Guo et al., 2019).

Definition 2.1. Consider a learning algorithm L and its unlearning function $U_L : (S, L(S), S_u) \to W^u$. The pair (L, U_L) achieves exact unlearning if $\forall S$ and $S_u \subset S$, $\mathsf{Pr}(L(S_r)) = \mathsf{Pr}(U_L(S, L(S), S_u))$, where $S_r = S \setminus S_u$. This implies that it becomes indistinguishable whether the model was trained after unlearning S_u from $L(S)$ or if it was trained exclusively on S_r. The pair (L, U_L) satisfies (γ, ζ)-unlearning if $\forall S$, $S_u \subset S$, and $\forall \mathcal{Z} \subseteq \mathcal{W}$, $\mathsf{Pr}(U_L(S, L(S), S_u) \in \mathcal{Z}) \leq e^\gamma \mathsf{Pr}(L(S_r) \in \mathcal{Z}) + \zeta$ and $\mathsf{Pr}(L(S_r) \in \mathcal{Z}) \leq e^\gamma \mathsf{Pr}(U_L(S, L(S), S_u) \in \mathcal{Z}) + \zeta$.

Related Work. Since the discovery of adversarial examples (Goodfellow et al., 2014), constructing adversarially robust models has become one of the most studied research topics (Mao et al., 2023; Glöckler et al., 2023; Attias & Hanneke, 2022; Ilyas et al., 2019; Schmidt et al., 2018; Sinha et al., 2023; Huai et al., 2022b). Among various existing defense strategies, adversarial training has been found to be the most effective approach against adversarial attacks (Mao et al., 2023; Singh et al., 2023; Yu et al., 2022a; Ghamizi et al., 2023; Wang & Wang, 2022; Qin et al., 2019; Goodfellow et al., 2014). Additionally, there is another prominent category of defense methodologies known as certified defense. These methods usually provide theoretically guaranteed bounds on the model's adversarial robustness (Zhang et al., 2022a; 2021; Gowal et al., 2018; Huai et al., 2022a; Yoshida & Miyato, 2017). For example, Gowal et al. (2018) employs the interval bound propagation to achieve fast and stable verified robust training. On the other hand, despite the great importance of studying adversarial robustness from the unlearning perspective, there is no existing work exploring the adversarial robustness properties of the unlearned models in the context of the right to be forgotten. Our work is different from Liu et al. (2023a), which investigates unlearning methods for adversarially trained models. Although recent works (Hu et al., 2023; Qian et al., 2023; Zhao et al., 2023; Di et al., 2022; Liu et al., 2024a) study potential pitfalls during unlearning, they fail to study adversarial robustness. Their main emphasis is on crafting malicious perturbations to perform data poisoning attacks. To the best of our knowledge, no prior research has examined the adversarial risks associated with the standard unlearning process from the perspective of adversarial attacks.

3. Adversarial Unlearning Attack

3.1. Building Motivation

From Definition 2.1, we know that machine unlearning provides a solution to mitigate these privacy risks, and involves the design of sophisticated unlearning techniques to remove private information from a trained machine learning model. Despite the focus on safeguarding individuals' private and sensitive data, existing unlearning methods neglect to assess the vulnerability of the unlearned models to adversarial attacks and security breaches. This oversight raises concerns regarding the adversarial robustness and overall security of the unlearning process. As no previous literature has studied the adversarial vulnerabilities of the unlearning system, for the ***threat model***, we start by discussing the adversary's objectives, capabilities, and knowledge for our attack. The objective of the adversary is to make malicious unlearning requests to deliberately undermine the adversarial robustness of the unlearned model. The adversary can generate the unlearning requests during the unlearning process. Since the AdvUA does not make any perturbations on the training data, in our main evaluation, we do not require a constraint for a malicious unlearning request as long as it is a training sample. We first consider a white-box setting where the AdvUA adversary knows the original model W^*, and subsequently, we delve into the black-box setting. Note that the white-box threat model represents the most powerful adversary that can appear in real-world scenarios and is of crucial importance to thoroughly study the adversary's behaviors.

(a) Low-density (b) High-density

Figure 2. Illustrations of the target test example and training examples in low-density region and high-density region.

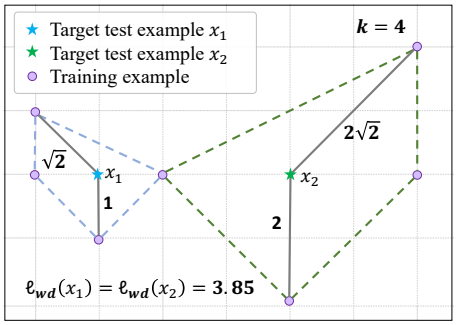


Figure 3. An illustration example of distance evaluation metric.

In the context of the right to be forgotten, *adversarial unlearning robustness* describes the property of an unlearned model to consistently predict the target class label for all perturbed inputs x' in an l_p-norm ball $\mathcal{B}_p^{\epsilon_p}(x) = \{x' \in \mathcal{X} : ||x - x'||_p \leq \epsilon_p\}$ of radius ϵ_p, after deleting the requested unlearning samples. To formalize this concept, we provide the following definition of adversarial robustness within the context of the right to be forgotten.

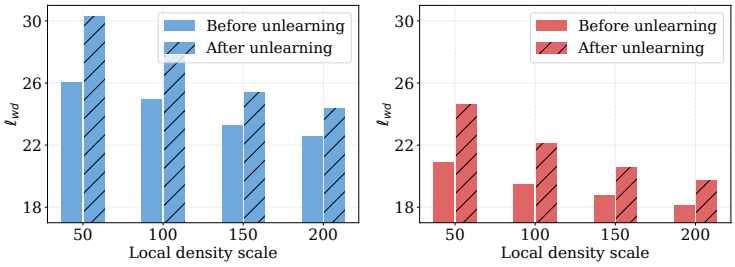
Definition 3.1 ($A_{S_u}^{\epsilon_p}(x)$ – Adversarial Unlearning Robustness)**.** Consider a sample $x \in \mathcal{X}$, a scalar ϵ_p, and a distance metric $\mathcal{D}(x, x') = ||x - x'||_p$. We use $F(x; W^u)$ to denote the unlearned model, which is derived by removing the requested unlearning samples S_u from the original model W^*. The unlearned model $F(x; W^u)$ is robust to adversarial perturbations of magnitude ϵ_p at input x if and only if $\forall(x + \delta) \in \mathcal{B}_p^{\epsilon_p}(x)$, $\arg\max_{c \in \mathcal{Y}} F(x; W^u)[c] = \arg\max_{c \in \mathcal{Y}} F(x+\delta; W^u)[c]$, where $W^u = U_L(S, L(S), S_u)$, and $\delta \in \mathbb{R}^d$ is the adversarial perturbation for x.

Our goal in this paper is to investigate whether the process of unlearning hinders an unlearned model's ability to withstand adversarial perturbations. Specifically, in this paper, we motivate our study with the previously raised scientific questions (i.e., *Q1* and *Q2* in Section 1). In Definition 3.1, we allow a bounded worst-case perturbation before passing the perturbed sample to the unlearned model W^u in the context of the right to be forgotten. In particular, based on the above definition, we can obtain that when $\epsilon_p = 0$ and $\arg\max_{c \in \mathcal{Y}} F(x; W^u)[c] \neq \arg\max_{c \in \mathcal{Y}} F(x+\delta; W^u)[c]$, *the adversary misleads the unlearned model W^u to directly misclassify the target victim samples without any further perturbations*. Without loss of generality, we set $p = \infty$ in this paper and omit this subscript for simplicity of notations. Additionally, using Definition 3.1, we can easily estimate the adversarial unlearning robustness of the unlearned model W^u across a population of data samples.

3.2. Formulation of AdvUA

As previously discussed (see Figure 1), we have observed that the adversarial robustness of unlearned models can be substantially compromised even with random data removal. To gain a deeper understanding, we now examine how unlearning samples affect both successful and unsuccessful adversarial examples. Then we present two visual examples that demonstrate the successful adversarial example in Figure 2a and the unsuccessful adversarial example in Figure 2b after unlearning. Here, we consider the first-order based unlearning method. Note that the two target test examples in Figure 2 cannot be adversarially attacked with the same perturbation sets *before unlearning*. Here we utilize the last representation layer and apply UMAP (McInnes et al., 2018) to project the adversarial examples, along with their 50 nearest neighbors. From this figure, we can observe that the target victim example shown in Figure 2a has now transitioned to a low-density region, and can easily fool the unlearned model. In contrast, the target victim example depicted in Figure 2b remains situated within a high-density region, making it difficult to generate a successful adversarial example for this particular target example.

Low-density Regions Generation. Drawing on the aforementioned observations, we make the following key contribution to exacerbating adversarial vulnerability in the unlearned models: employing unlearning techniques to strategically position target victim samples within low-density regions. As a result, the unlearned models may not have learned robust decision boundaries or patterns for these regions, making them more susceptible to adversarial attacks and misclassifications. Let $\mathcal{V} = \{(x_v, y_v)\}_{v=1}^{V}$ represent the victim samples that adversaries intend to attack. We use $\mathcal{N}_{K_v}(x_v)$ to denote the top-K_v nearest neighbors for the target sample x_v. Let $x_v^{max} \in \mathcal{N}_{K_v}(x_v)$ denote the sample that has the largest distance from the target victim sample x_v. To estimate the space-filling capability of the region surrounding sample x_v, we compute the relative distance information



(a) Success cases (b) Failure cases

Figure 4. Density of success and failure cases before and after unlearning measured by ℓ_{wd} on different local scales. A high ℓ_{wd} value indicates a low-density region.

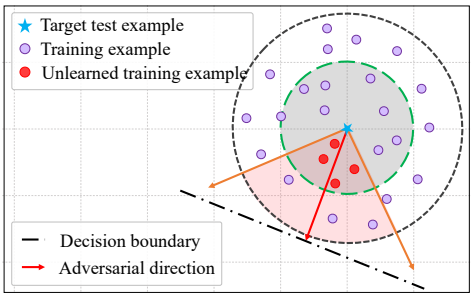


Figure 5. Illustration of adversarial direction alignment for unlearning training examples.

of its neighbors, with respect to x_v^{max}, utilizing classical expansion models (Ma et al., 2018; Houle, 2017; Houle et al., 2012; Karger & Ruhl, 2002). Specifically, for x_v, the relative distance of $x_i \in \mathcal{N}_{K_v}(x_v)$ with respect to x_v^{max} in the representation space is calculated as $\log(||H(x_i;W^u) - H(x_v;W^u)||_2 / ||H(x_v^{max};W^u) - H(x_v;W^u)||_2)$. However, if we directly calculate this distance value, we may encounter the case where $\forall x_i \in \mathcal{N}_{K_v}(x_v)$, $||H(x_i;W^u) - H(x_v;W^u)||_2 = ||H(x_v^{max};W^u) - H(x_v;W^u)||_2$. On the other hand, relying solely on relative distances can be challenging, as they do not provide insights into the absolute values of data points. As illustrated in Figure 3, even though x_1 and x_2 have the same relative distance values, x_2 is more likely to be vulnerable to adversarial attacks. Therefore, in addition to the relative distances, we should also take absolute distance information into account. For x_v and its neighbors $\mathcal{N}_{K_v}(x_v)$, we can calculate the Euclidean distance between this sample and its neighbor $x_i \in \mathcal{N}_{K_v}(x_v)$ as $D(H(x_i;W^u), H(x_v;W^u)) = ||H(x_i;W^u) - H(x_v;W^u)||_2^2$, where $H(x_i;W^u)$ is the representation of x_i. By combining the above, we formulate the following measure

$$\ell_{wd}(x_v, \mathcal{N}_{K_v}(x_v); W^u) = -\lambda_1 \Big(\frac{1}{K_v} \sum_{x_i \in \mathcal{N}_{K_v}(x_v)} \log \frac{||H(x_i;W^u) - H(x_v;W^u)||_2}{||H(x_v^{max};W^u) - H(x_v;W^u)||_2 + \beta}\Big)^{-1} + \frac{(1-\lambda_1)}{K_v} \sum_{x_i \in \mathcal{N}_{K_v}(x_v)} D(H(x_i;W^u), H(x_v;W^u)), \tag{1}$$

where $\lambda_1 \in [0,1]$ and $K_v = |\mathcal{N}_{K_v}(x_v)|$. In the above, β is a pre-defined value to avoid cases where the fraction has a zero denominator. The first term in ℓ_{wd} is used to quantify the local density around x_i measured in the space subsequent to unlearning S_u (Zhang et al., 2019; Ma et al., 2018; Houle, 2017). The second one ensures a comprehensive characterization of the local behavior around x_v.

To assess the impact of unlearning on density as expressed in Eqn. (1), we focus on *hard samples* that are attacked successfully (yielding incorrect predictions) in standard training but failed (resulting in correct predictions) in adversarial training. In Figure 4, we adopt the first-order based unlearning method and compare the average density of success cases (attack success after unlearning) and failure cases (attack fail after unlearning) within different local density scales (nearest neighbors). More details can be found in the full version of this paper. Figure 4 shows that success cases typically have higher ℓ_{wd} compared to failure cases after unlearning, which suggests that success cases are located in relatively low-density regions and are thus susceptible to attacks (consistent with observations in Figure 2).

Attack Direction Alignment. From the above, we know that for a target sample x_v located in a low-density region, it tends to experience successful attacks during inference since it is less likely to be covered by the distribution of the training samples. The next question we want to explore is whether all of the nearest samples are equally critical for performing adversarial unlearning attacks to assign a wrong label to x_v. However, we find an intriguing phenomenon: not all of the nearest samples are equally important for adversarial unlearning attacks. This phenomenon reveals that for x_v, the efficient unlearning directions towards the low-density regions should align well with its adversarial attack direction, which is determined by $\nabla_{x_v}\mathcal{L}(x_v, y_v; W)$, where y_v is the label of x_v. To help understanding, we give an illustration example for the attack direction alignment in Figure 5, where the grey area represents the high-density region for the target test example and red zones highlight directions that are positively aligned with the red arrow. From this figure, the samples that fall within the intersection of the grey and red areas are the ones we should focus on. Given sample x_v and its neighbor sample $x_i \in \mathcal{N}_{K_v}(x_v)$, we propose the following measure to estimate the alignment between the unlearning attack and the adversarial attack

$$\ell_{dirc}(x_i, x_v; W^u) = \frac{(H(x_i;W^u) - H(x_v;W^u)) \cdot (\nabla_H \mathcal{L}(x_v, y_v; W^u))}{||H(x_i;W^u) - H(x_v;W^u)||_2 * ||\nabla_H \mathcal{L}(x_v, y_v; W^u)||_2}, \tag{2}$$

where $H(x_i; W^u) - H(x_v; W^u)$ is the unlearning direction when we unlearn a sample $x_i \in \mathcal{N}_{K_v}(x_v)$. In the above, we use the normalized adversarial attack to remove the influence of the scaling factor when comparing different models. Based on the equation, for the effective unlearning samples, they should closely match the adversarial attack with the unlearning direction.

Overall Loss. Based on the above observations, to understand the worst-case attack performance, we design a novel adversarial unlearning attack framework to increase the inherent adversarial vulnerability of the unlearned models. Let $S_e = \{x_t\}_{t=1}^T$ represent a subset of S that is accessible to the adversary. For each $x_t \in S_e$, we define a discrete indication parameter $\xi_t \in \{0, 1\}$ to indicate whether the sample x_t should be completely deleted ($\xi_t = 1$) or not ($\xi_t = 0$). The forget set S_u to be unlearned is denoted as $S_u = S_e \circ \Phi = \{x_t | x_t \in S_e \text{ and } \xi_t = 1\}$, where $\Phi = \{\xi_t\}_{t=1}^T$. We use $\{(x_v, y_v)\}_{v=1}^V$ to denote the *hard* target samples, which cannot be successfully attacked from the original model W^* using the same perturbation budget ϵ. Therefore, to effectively attack these hard target samples (i.e., $\{(x_v, y_v)\}_{v=1}^V$), we formulate the below overall loss

$$\begin{aligned} &\max_{\{\delta_v \in \mathcal{B}^\epsilon(x_v)\}_{v=1}^V} \sum_{v=1}^V \mathcal{L}(x_v + \delta_v, y_v; W^u(\Phi)) \\ s.t. \Phi \leftarrow \arg\max_{\Phi} &\sum_{v=1}^V \ell_{wd}(x_v, \mathcal{N}_{K_v}(x_v); W^u(\Phi)) \\ &- \sum_{v=1}^V \frac{\lambda_2}{K_v} \sum_{x_i \in \mathcal{N}_{K_v}(x_v)} \ell_{dirc}(x_i, x_v; W^u(\Phi)), \end{aligned} \tag{3}$$

where $K_v = |\mathcal{N}_{K_v}(x_v)|$ and $W^u(\Phi) = U_L(S, W^*, S_u = S_e \circ \Phi)$. $\mathcal{L}$ is the loss function to enforce the adversarial example $x_v + \delta_v$ to be predicted as a different label than y_v. Notably, in the above, when $\delta_v = 0$, the requested malicious unlearning samples can cause the unlearned model W^u to directly misclassify sample x_v without the need for any further adversarial manipulations. Exactly solving Eqn. (3) would be computationally infeasible (Korte et al., 2011), and instead, we refer to an empirical greedy approach (Barron et al., 2008) and a discrete relaxation approach to solve the above optimization problem. Due to space limitation, both the algorithms for optimizing the above formulated overall loss and the corresponding computational complexity analysis are deferred to the full version of this paper.

Theorem 3.2. *Consider a data distribution $\mathcal{X}$ characterized by a Gaussian distribution with mean $\mu \in \mathbb{R}^d$ and variance $\sigma^2 I$, i.e., $\mathcal{X} \sim N(\mu, \sigma^2 I)$. Let $\{x_i\}_{i=1}^n$ be a set of samples drawn from $N(\mu, \sigma^2 I)$. Then the expected local density around point $\tilde{x}$ is lower bounded by*

$$\begin{aligned} &\mathbb{E}_{\{x_i\}_{i=1}^n \sim \mathcal{X}} \left[\sum_{i=1}^n \mathbb{1}\left\{ \|x_i - \tilde{x}\|_2^2 \le q \right\} \right] \ge \\ &n \times \left[1 - \frac{\sigma^2 d}{\left(q - \|\mu - \tilde{x}\|_2^2\right)^2} \right] \end{aligned} \tag{4}$$

where $\tilde{x} \in \mathbb{R}^d$ and $q \in \mathbb{R}$.

Proof. To begin with, the expectation $\mathbb{E}_{\{x_i\}_{i=1}^n \sim \mathcal{X}} \left[\sum_{i=1}^n \mathbb{1}\left\{ \|x_i - \tilde{x}\|_2^2 \le k\sigma \right\} \right]$ is expressed as the expected value of a binomial distribution with N trials. Then we can have

$$\begin{aligned} &\mathbb{E}_{\{x_i\}_{i=1}^n \sim \mathcal{X}} \left[\sum_{i=1}^n \mathbb{1}\left\{ \|x_i - \tilde{x}\|_2^2 \le k\sigma \right\} \right] \\ =& \sum_{i=1}^n \mathbb{E}_{x_i \sim \mathcal{X}} \left[\mathbb{1}\left\{ \|x_i - \tilde{x}\|_2^2 \le k\sigma \right\} \right] \\ =& \; n \times P\left(\|x_i - \tilde{x}\|_2^2 \le k\sigma \right) \\ =& \; n \times P\left(\|x_i - \mu + \mu - \tilde{x}\|_2^2 \le k\sigma \right) \\ \ge& \; n \times P\left(\|x_i - \mu\|_2^2 + \|\mu - \tilde{x}\|_2^2 \le k\sigma \right) \\ =& \; n \times P\left(\|x_i - \mu\|_2^2 \le k\sigma - \|\mu - \tilde{x}\|_2^2 \right) \\ =& \; n \times P\left(\|x_i - \mu\|_2^2 \le \sigma \cdot \left(k - \frac{1}{\sigma}\|\mu - \tilde{x}\|_2^2 \right) \right) \\ =& \; n \times P\left(\sqrt{(x_i - \mu)^\top \sigma^{-2} I \, (x_i - \mu_i)} \right. \\ & \qquad \left. \le k - \frac{1}{\sigma}\|\mu - \tilde{x}\|_2^2 \right). \end{aligned} \tag{5}$$

Upon applying the Multidimensional Chebyshev's Inequality (Chen, 2007), we can obtain

$$\begin{aligned} &\mathbb{E}_{\{x_i\}_{i=1}^n \sim \mathcal{X}} \left[\sum_{i=1}^n \mathbb{1}\left\{ \|x_i - \tilde{x}\|_2^2 \le k\sigma \right\} \right] \\ &\ge n \times \left[1 - \frac{d}{\left(k - \frac{1}{\sigma}\|\mu - \tilde{x}\|_2^2\right)^2} \right]. \end{aligned} \tag{6}$$

Let $q = k\sigma$, then we can get

$$\begin{aligned} &\mathbb{E}_{\{x_i\}_{i=1}^n \sim \mathcal{X}} \left[\sum_{i=1}^n \mathbb{1}\left\{ \|x_i - \tilde{x}\|_2^2 \le q \right\} \right] \\ &\ge n \times \left[1 - \frac{\sigma^2 d}{\left(q - \|\mu - \tilde{x}\|_2^2\right)^2} \right]. \end{aligned} \tag{7}$$

□

Theorem 3.3. *Let g_n be any learning algorithm, i.e., a function from $n \geq 0$ samples in $\mathbb{R}^d \times \{\pm 1\}$ to a binary classifier f_n. Moreover, let $W \in \mathbb{R}^d$ be the weight of f_n and $W = \frac{1}{n}\sum_{i=1}^{n} y_i x_i$, and let $\theta \in \mathbb{R}^d$ be drawn from $N(0, I)$, $\|\theta\|_2 = \sqrt{d}$. We draw n_1, n_2 samples from the (θ, σ)-Gaussian model, which generates $(x, y) \in \mathbb{R}^d \times \{\pm 1\}$ by first randomly selecting a label $y \in \{\pm 1\}$ and then sampling $x \in \mathbb{R}^d$ from $N(y \cdot \theta^\star, \sigma^2 I)$. Let the expected l_∞^ϵ-robust classification errors of f_{n_1}, f_{n_2} are R_1, R_2. Then it can be deduced that $R_1 \leq R_2$ holds with a probability at least $1 - 2\exp\left(-\frac{d}{8(\sigma^2+1)}\right)$ if $n_1 \geq c_2\epsilon^2\sqrt{d}$ and $n_2 \leq \frac{\epsilon^2\sigma^2}{8\log d}$, where $0 \leq \sigma \leq c_1 d^{1/4}$, $\sqrt{\frac{8\log d}{\sigma^2}} \leq \epsilon \leq \frac{1}{2}$.*

In Theorem 3.2, we estimate the local density at a given point $\tilde{x}$ by counting the number of data covered within a ball centered as $\tilde{x}$ with radius q. This theorem demonstrates that as the quantity of training data decreases, the average local density also experiences a proportional decline. Theorem 3.3 illustrates the relation between the number of training samples and the associated l_∞^ϵ-adversarial robust error. Theorem 3.3 shows that when the training set size diminishes from n_1 to n_2 under the conditions mentioned in this theorem, there is a high probability that the robust error will increase. The proof of Theorem 3.3 is deferred to the full version of this paper.

Discussions on the Black-box Setting. In the black-box scenario, it's assumed that the adversaries are unaware of any prior knowledge about the target well-trained model, including the model architecture and model parameters. Following existing works (Wen et al., 2023; Liu et al., 2023b; Byun et al., 2022; Di et al., 2022; Zhou et al., 2018; Liu et al., 2017), the black-box attack can be executed by constructing a substitute model and transferring adversarial examples, which leverages shared decision boundaries among various models. This approach can reduce exposure risk, and hence, many existing works have been focused on adversarial transferability. In our black-box setting, we can train several substitute models and transfer both the selected unlearning samples and generated adversarial examples to attack the target victim model.

Enhanced Model Stealing Attacks. Traditional model stealing attacks (Genç et al., 2023; Yu et al., 2020; Tramèr et al., 2016) first send a series of queries Q to the target victim model and then use the collected query-response pairs to train a surrogate model that approximates the behavior of the victim model. The key to the success of model stealing is to find query samples lying approximately on the decision boundary of the victim model, which is not easy. To improve the query effectiveness, our goal here is to make malicious unlearning requests to learn the decision boundary information of the victim target model W^*. Instead of starting from scratch, following existing works, we first construct an initialized primitive surrogate model. Here, we consider the scenarios where the adversary has a set of query samples Q for querying the victim target model. In the context of machine unlearning, we also assume the adversary has available unlearning samples, denoted as S_e. Based on AdvUA (in Eqn. (3)), we can select a set of malicious unlearning samples to increase its query effectiveness by revealing the important decision boundary information of the victim model.

4. Experimental Results

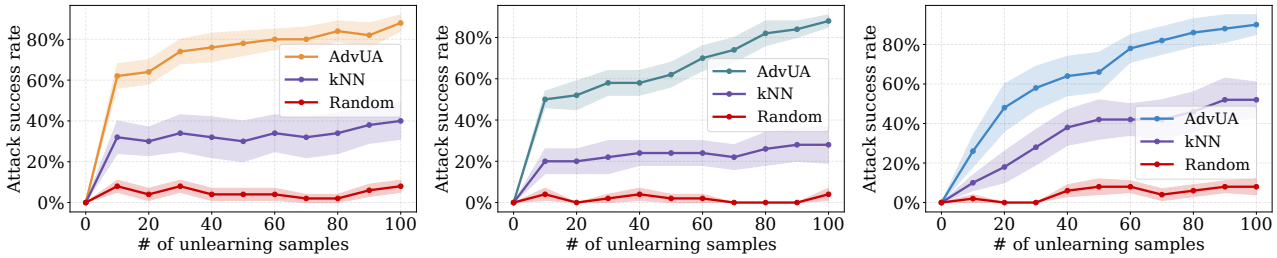
In this section, we conduct experiments to validate the performance of AdvUA. All experiments are performed for 10 trials, and we report the mean and standard errors in the following analyses. For more experimental details (e.g., experimental setup and parameter settings) and experimental results (e.g., running time, attack performance against certified defenses, more unlearning methods and ablation studies), please refer to the full version of this paper.

4.1. Experimental Setup

Datasets and Models. In experiments, we adopt the following datasets: ImageNet (Deng et al., 2009), CIFAR-10 (Krizhevsky & Hinton, 2010), SVHN (Netzer et al., 2011), and IRIS (Fisher, 1988). We consider various deep learning models, including ResNet-50, ResNet-18 (He et al., 2016), DenseNet-121 (Huang et al., 2017), VGG-19 (Simonyan & Zisserman, 2015), a 5-layer ConvNet with max-pooling and dropout, and a multilayer perceptron (MLP).

Baselines. We compare the performance of AdvUA with the following baselines: *random* deletion and *k-nearest neighbors (kNN)* deletion. The random deletion method randomly removes some training samples from the training set, regardless of the target sample. The kNN deletion method removes the k closest training samples to a target sample in the input space.

Implementation Details. We conduct comprehensive experiments to evaluate the attack performance of AdvUA on both undefended and defended models. The undefended models are constructed using natural training algorithms, and the defended models are constructed using adversarial training and certified defense methods. The adversarial training involves training the model against a PGD adversary with l_∞ project gradient descent of $\epsilon = 8/255$ (Madry et al., 2018). The certified defense methods utilize techniques including spectral norm regularization (Yoshida & Miyato, 2017) and interval bound propagation (Gowal et al., 2018) to provide provably robustness bounds. We evaluate the robustness of undefended and defended models against FGSM (Goodfellow et al., 2015), PGD (Madry et al., 2018), and CW (Carlini & Wagner, 2017) adversarial attacks. Specifically, we



(a) FGSM (b) PGD (c) CW

Figure 6. Attack performance of AdvUA on adversarially trained models against various attacks.

use a perturbation budget of $\epsilon = 8/255$ to generate adversarial examples, setting the iteration steps to 7 for PGD attacks and 30 for CW attacks. Regarding the unlearning methods for removing the training samples, we select the first-order based unlearning method (Warnecke et al., 2023), the second-order based unlearning method (Warnecke et al., 2023), the unrolling SGD unlearning method (Thudi et al., 2022), and SISA (Bourtoule et al., 2021a).

Table 1. Attack transferability of FGSM with AdvUA.

Method	# of unlearning samples	VGG-19	DenseNet-121
Baseline	None	$63.0\% \pm 4.7\%$	$61.0\% \pm 6.9\%$
AdvUA	10	$\mathbf{88.0\% \pm 2.9\%}$	$\mathbf{88.0\% \pm 2.9\%}$
	20	$\mathbf{88.0\% \pm 2.9\%}$	$\mathbf{90.0\% \pm 1.5\%}$
	30	$\mathbf{92.0\% \pm 2.0\%}$	$\mathbf{96.0\% \pm 1.6\%}$

4.2. Attack Performance Against Robust Training

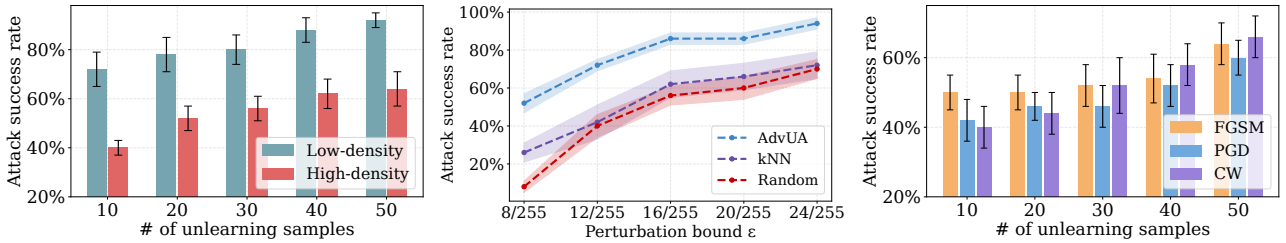
We first investigate the attack performance of AdvUA on the defended models against FGSM, PGD, and CW attacks. The defended models are adversarially trained with ResNet-18 on CIFAR-10 and ConvNet on SVHN. The results are shown in Figure 6a and Figure 6b for CIFAR-10, and Figure 6c for SVHN. Here, we randomly select hard samples that are correctly classified on the defended model and then remove various quantities of training samples for each sample using the first-order based unlearning method. As shown in the figures, AdvUA significantly enhances the attack success rates on the defended models. The increase in vulnerability can be primarily because, after the unlearning process, test samples shift to low-density regions that adversarially trained models struggle to cover effectively, especially along the adversarial direction. As more training samples are removed, these test samples migrate to even sparser regions, resulting in higher attack success rates. For instance, when unlearning 100 training samples for each hard sample, AdvUA achieves approximately 88% attack success rates on CIFAR-10 against FGSM and PGD attacks and 90% on SVHN against CW attacks. In comparison, the attack performance of the random baseline is notably poor. While the kNN baseline slightly improves, it still significantly lags behind AdvUA. One reason is that although the kNN might reduce local density to a degree, it is not as effective as AdvUA. Additionally, AdvUA uniquely aligns the direction of unlearning with the adversarial direction. Therefore, this experiment illustrates the efficacy and superiority of AdvUA in selecting crucial unlearning samples to achieve the desired attack goals.

Table 2. Test accuracy of model stealing attacks.

Query count	Query-based attack	**Query-based attack + unlearn (AdvUA)**
100	$45.24\% \pm 6.17\%$	$\mathbf{85.71\% \pm 4.92\%}$ $\mathbf{(1.89\times)}$
200	$56.19\% \pm 9.85\%$	$\mathbf{88.57\% \pm 2.41\%}$ $\mathbf{(1.58\times)}$
300	$85.56\% \pm 2.81\%$	$\mathbf{90.00\% \pm 2.85\%}$ $\mathbf{(1.05\times)}$
400	$86.11\% \pm 5.74\%$	$\mathbf{92.22\% \pm 1.41\%}$ $\mathbf{(1.07\times)}$

4.3. Attack Transferability and Model Stealing Attacks

Effectiveness on Attack Transferability. We consider a black-box scenario to examine the effectiveness of AdvUA on adversarial transferability. We use a substitute model to unlearn various numbers of training samples with the first-order based method and then generate the adversarial examples. Subsequently, these unlearning samples and adversarial examples are transferred to attack the target model. Table 1 presents the attack transferability of FGSM on the undefended models on ImageNet, where we adopt the ResNet-50 as the substitute model and the VGG-19 and DenseNet-121 as the target models. For comparison, we also include the attack transferability of the baseline method, which directly applies generated adversarial samples to attack the target models (without unlearning). Conversely, AdvUA employs unlearning to create a sparse local density environment around the test sample prior to the transfer process. The results show that AdvUA outperforms the baseline by a large margin in both target models. For instance, when transferring ResNet-50 to DenseNet-121, AdvUA achieves a 96% attack success rate with 30 unlearning samples, while



(a) Sampling density (b) Perturbation bound (c) Unlearning method

Figure 7. Ablation studies over sampling density, perturbation bound, and unlearning method.

the baseline only receives a 61% attack success rate. These findings indicate that AdvUA effectively boosts attack transferability in the realm of adversarial machine learning.

Effectiveness on Model Stealing Attacks. We also evaluate the effectiveness of AdvUA on model stealing attacks. We initially train a target model composed of an MLP using the IRIS dataset and then employ a synthetic dataset as queries to steal this model, as outlined in Tramèr et al. (2016). Table 2 shows the test accuracy of the extracted model obtained with and without unlearning. Regarding the unlearning method, we utilize the first-order based approach. The numbers in the parentheses represent the improvement of test accuracy by integrating AdvUA. The results indicate that our method boosts the performance of model stealing attacks, especially when queries are limited. For instance, AdvUA enhances the test accuracy of the extracted model by a factor of $1.89\times$ with 100 queries compared to the original model stealing attacks. By strategically unlearning samples to diminish the local density near the target query, especially close to the decision boundary, we create a localized region with reduced robustness around the target query. This adjustment makes the query more effective in training the derived model, leading to superior accuracy compared to the original query-based attack. In summary, AdvUA effectively improves the performance of model stealing attacks, providing another perspective to validate the influence of unlearning on model robustness. The corresponding query time can be found in the full version of this paper.

4.4. Ablation Study

In this section, we conduct ablation studies over sampling density, perturbation bound, and unlearning methods. We adopt the first-order based unlearning method for Figure 7a and Figure 7b, and the SISA unlearning method for Figure 7c. As depicted in Figure 7a, when unlearning the same number of training samples on CIFAR-10, test samples initially from low-density regions exhibit higher PGD attack success rates than those from high-density regions. Notably, even when selecting test samples from high-density regions, AdvUA still achieves remarkable outcomes. We also compare attack success rates with different perturbation bounds during CW attacks on CIFAR-10. As shown in Figure 7b, despite larger perturbation bounds benefiting baselines, AdvUA consistently outperforms them. In addition, Figure 7c illustrates the effectiveness of AdvUA with the SISA unlearning method, to attack the defended model against FGSM, PGD, and CW attacks on CIFAR-10. For example, AdvUA can achieve around 66% CW attack success rates with 50 unlearning samples. In these ablation studies, our experimental results emphasize the efficacy of AdvUA in increasing the defended model's vulnerability, irrespective of variations in sampling density, perturbation bounds, and unlearning techniques.

5. Conclusion and Future Work

In this paper, we take a significant step towards developing a comprehensive understanding of the adversarial risks associated with machine unlearning. We show that the low density of the target sample in the space, along with the alignment between adversarial attacks and unlearning directions, are crucial factors for generating successful adversarial examples, which were not achievable prior to unlearning. Drawing upon our insightful observations, in this paper, we design a new adversarial unlearning attack (AdvUA), with the ultimate goal of exacerbating the adversarial vulnerability of the unlearned models. Additionally, AdvUA can also make model stealing attacks more effective and stealthy. Our extensive experimental results serve as strong empirical evidence of the effectiveness and computational efficiency of the proposed attack, underscoring the importance of considering security implications alongside data privacy concerns within the domain of machine unlearning.

In the future, we will investigate the detection and defense mechanisms to mitigate and defend against adversarial unlearning attacks in the context of the right to be forgotten. Besides deep learning models, we will also investigate the potential threats of adversarial unlearning attacks in other domains (e.g., federated learning, and graph neural networks) using different unlearning methods.

Impact Statement

In this paper, we uncover a significant adversarial threat inherent within the machine unlearning process, wherein the model's adversarial robustness may suffer a notable decline. Addressing this issue is critical for the development of reliable and robust unlearning systems, particularly in light of the escalating demand for privacy and data protection.

References

Attias, I. and Hanneke, S. Adversarially robust learning of real-valued functions. *arXiv preprint arXiv:2206.12977*, 2022.

Bai, T., Luo, J., Zhao, J., Wen, B., and Wang, Q. Recent advances in adversarial training for adversarial robustness. *arXiv preprint arXiv:2102.01356*, 2021.

Barron, A. R., Cohen, A., Dahmen, W., and DeVore, R. A. Approximation and learning by greedy algorithms. *The annals of statistics*, 36(1):64–94, 2008.

Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C., Jia, H., Travers, A., Zhang, B., Lie, D., and Papernot, N. Machine unlearning. In *Proceedings of the 42nd IEEE Symposium on Security and Privacy*, May 2021a.

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

Brophy, J. and Lowd, D. Machine unlearning for random forests. In *International Conference on Machine Learning*, pp. 1092–1104. PMLR, 2021.

Byun, J., Cho, S., Kwon, M.-J., Kim, H.-S., and Kim, C. Improving the transferability of targeted adversarial examples through object-based diverse input. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15244–15253, June 2022.

Carlini, N. and Wagner, D. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pp. 39–57. Ieee, 2017.

Che, T., Zhou, Y., Zhang, Z., Lyu, L., Liu, J., Yan, D., Dou, D., and Huan, J. Fast federated machine unlearning with nonlinear functional theory. In *International conference on machine learning*, pp. 4241–4268. PMLR, 2023.

Chen, M., Zhang, Z., Wang, T., Backes, M., Humbert, M., and Zhang, Y. When machine unlearning jeopardizes privacy. In *Proceedings of the 2021 ACM SIGSAC conference on computer and communications security*, pp. 896–911, 2021.

Chen, M., Zhang, Z., Wang, T., Backes, M., Humbert, M., and Zhang, Y. Graph unlearning. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pp. 499–513, 2022.

Chen, X. A new generalization of chebyshev inequality for random vectors. *arXiv preprint arXiv:0707.0805*, 2007.

Chourasia, R. and Shah, N. Forget unlearning: Towards true data-deletion in machine learning. In *International Conference on Machine Learning*, pp. 6028–6073. PMLR, 2023.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Di, J. Z., Douglas, J., Acharya, J., Kamath, G., and Sekhari, A. Hidden poison: Machine unlearning enables camouflaged poisoning attacks. In *NeurIPS ML Safety Workshop*, 2022.

Fisher, R. A. Iris. UCI Machine Learning Repository, 1988. DOI: https://doi.org/10.24432/C56C76.

Genç, D., Özuysal, M., and Tomur, E. A taxonomic survey of model extraction attacks. In *2023 IEEE International Conference on Cyber Security and Resilience (CSR)*, pp. 200–205. IEEE, 2023.

Ghamizi, S., Zhang, J., Cordy, M., Papadakis, M., Sugiyama, M., and Le Traon, Y. Gat: guided adversarial training with pareto-optimal auxiliary tasks. In *International Conference on Machine Learning*, pp. 11255–11282. PMLR, 2023.

Glöckler, M., Deistler, M., and Macke, J. H. Adversarial robustness of amortized bayesian inference. *arXiv preprint arXiv:2305.14984*, 2023.

Goodfellow, I., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Gowal, S., Dvijotham, K., Stanforth, R., Bunel, R., Qin, C., Uesato, J., Arandjelovic, R., Mann, T., and Kohli, P. On the effectiveness of interval bound propagation for training verifiably robust models. *arXiv preprint arXiv:1810.12715*, 2018.

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

Guo, C., Goldstein, T., Hannun, A., and Van Der Maaten, L. Certified data removal from machine learning models. *arXiv preprint arXiv:1911.03030*, 2019.

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

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Houle, M. E. Local intrinsic dimensionality i: an extreme-value-theoretic foundation for similarity applications. In *Similarity Search and Applications: 10th International Conference, SISAP 2017, Munich, Germany, October 4-6, 2017, Proceedings 10*, pp. 64–79. Springer, 2017.

Houle, M. E., Kashima, H., and Nett, M. Generalized expansion dimension. In *2012 IEEE 12th International Conference on Data Mining Workshops*, pp. 587–594. IEEE, 2012.

Hu, H., Wang, S., Chang, J., Zhong, H., Sun, R., Hao, S., Zhu, H., and Xue, M. A duty to forget, a right to be assured? exposing vulnerabilities in machine unlearning services. *arXiv preprint arXiv:2309.08230*, 2023.

Huai, M., Liu, J., Miao, C., Yao, L., and Zhang, A. Towards automating model explanations with certified robustness guarantees. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 6935–6943, 2022a.

Huai, M., Zheng, T., Miao, C., Yao, L., and Zhang, A. On the robustness of metric learning: an adversarial perspective. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 16(5):1–25, 2022b.

Huang, G., Liu, Z., Van Der Maaten, L., and Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., and Madry, A. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

Jia, J., Liu, J., Ram, P., Yao, Y., Liu, G., Liu, Y., Sharma, P., and Liu, S. Model sparsification can simplify machine unlearning. *arXiv preprint arXiv:2304.04934*, 2023.

Karger, D. R. and Ruhl, M. Finding nearest neighbors in growth-restricted metrics. In *Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pp. 741–750, 2002.

Korte, B. H., Vygen, J., Korte, B., and Vygen, J. *Combinatorial optimization*, volume 1. Springer, 2011.

Krizhevsky, A. and Hinton, G. Convolutional deep belief networks on cifar-10. *Unpublished manuscript*, 40(7): 1–9, 2010.

Li, B., Chen, C., Wang, W., and Carin, L. Certified adversarial robustness with additive noise. *Advances in neural information processing systems*, 32, 2019.

Liu, J., Xue, M., Lou, J., Zhang, X., Xiong, L., and Qin, Z. Muter: Machine unlearning on adversarially trained models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4892–4902, 2023a.

Liu, Y., Chen, X., Liu, C., and Song, D. Delving into transferable adversarial examples and black-box attacks. In *International Conference on Learning Representations*, 2017.

Liu, Z., Zhao, Z., and Larson, M. Image shortcut squeezing: Countering perturbative availability poisons with compression. In *International Conference on Machine Learning*, 2023b. URL https://api.semanticscholar.org/CorpusID:256416324.

Liu, Z., Wang, T., Huai, M., and Miao, C. Backdoor attacks via machine unlearning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 14115–14123, 2024a.

Liu, Z., Ye, H., Chen, C., and Lam, K.-Y. Threats, attacks, and defenses in machine unlearning: A survey. *arXiv preprint arXiv:2403.13682*, 2024b.

Ma, X., Li, B., Wang, Y., Erfani, S. M., Wijewickrema, S., Schoenebeck, G., Song, D., Houle, M. E., and Bailey, J. Characterizing adversarial subspaces using local intrinsic dimensionality. In *International Conference on Learning Representations*, 2018.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Maini, P., Wong, E., and Kolter, Z. Adversarial robustness against the union of multiple perturbation models. In *International Conference on Machine Learning*, pp. 6640–6650. PMLR, 2020.

Mao, C., Geng, S., Yang, J., Wang, X., and Vondrick, C. Understanding zero-shot adversarial robustness for large-scale models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=P4bXCawRi5J.

McInnes, L., Healy, J., and Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Neel, S., Roth, A., and Sharifi-Malvajerdi, S. Descent-to-delete: Gradient-based methods for machine unlearning. In *Algorithmic Learning Theory*, pp. 931–962. PMLR, 2021.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. 2011.

Nguyen, T. T., Huynh, T. T., Nguyen, P. L., Liew, A. W.-C., Yin, H., and Nguyen, Q. V. H. A survey of machine unlearning. *arXiv preprint arXiv:2209.02299*, 2022.

Pang, T., Yang, X., Dong, Y., Su, H., and Zhu, J. Bag of tricks for adversarial training. *arXiv preprint arXiv:2010.00467*, 2020.

Qian, W., Zhao, C., Shao, H., Chen, M., Wang, F., and Huai, M. Patient similarity learning with selective forgetting. In *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 529–534. IEEE, 2022.

Qian, W., Zhao, C., Le, W., Ma, M., and Huai, M. Towards understanding and enhancing robustness of deep learning models against malicious unlearning attacks. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1932–1942, 2023.

Qin, C., Martens, J., Gowal, S., Krishnan, D., Dvijotham, K., Fawzi, A., De, S., Stanforth, R., and Kohli, P. Adversarial robustness through local linearization. *Advances in Neural Information Processing Systems*, 32, 2019.

Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., and Madry, A. Adversarially robust generalization requires more data. *Advances in neural information processing systems*, 31, 2018.

Sekhari, A., Acharya, J., Kamath, G., and Suresh, A. T. Remember what you want to forget: Algorithms for machine unlearning. *Advances in Neural Information Processing Systems*, 34:18075–18086, 2021.

Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

Singh, N. D., Croce, F., and Hein, M. Revisiting adversarial training for imagenet: Architectures, training and generalization across threat models. In *NeurIPS*, 2023.

Sinha, S., Huai, M., Sun, J., and Zhang, A. Understanding and enhancing robustness of concept-based models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 15127–15135, 2023.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Tarun, A. K., Chundawat, V. S., Mandal, M., and Kankanhalli, M. Deep regression unlearning. In *International Conference on Machine Learning*, pp. 33921–33939. PMLR, 2023.

Thudi, A., Deza, G., Chandrasekaran, V., and Papernot, N. Unrolling sgd: Understanding factors influencing machine unlearning. In *2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P)*, pp. 303–319. IEEE, 2022.

Tramèr, F., Zhang, F., Juels, A., Reiter, M. K., and Ristenpart, T. Stealing machine learning models via prediction {APIs}. In *25th USENIX security symposium (USENIX Security 16)*, pp. 601–618, 2016.

Voigt, P. and Von dem Bussche, A. The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, 10(3152676): 10–5555, 2017.

Wang, C.-L., Huai, M., and Wang, D. Inductive graph unlearning. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 3205–3222, 2023.

Wang, H. and Wang, Y. Self-ensemble adversarial training for improved robustness. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=oU3aTsmeRQV.

Warnecke, A., Pirch, L., Wressnegger, C., and Rieck, K. Machine unlearning of features and labels. *Network and Distributed System Security Symposium*, 2023.

Wen, R., Zhao, Z., Liu, Z., Backes, M., Wang, T., and Zhang, Y. Is adversarial training really a silver bullet for mitigating data poisoning? In *The Twelfth International Conference on Learning Representations*, 2023.

Yan, H., Li, X., Guo, Z., Li, H., Li, F., and Lin, X. Arcane: An efficient architecture for exact machine unlearning. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pp. 4006–4013, 2022.

Yoshida, Y. and Miyato, T. Spectral norm regularization for improving the generalizability of deep learning. *arXiv preprint arXiv:1705.10941*, 2017.

Yu, C., Han, B., Shen, L., Yu, J., Gong, C., Gong, M., and Liu, T. Understanding robust overfitting of adversarial

training and beyond. In *International Conference on Machine Learning*, pp. 25595–25610. PMLR, 2022a.

Yu, H., Yang, K., Zhang, T., Tsai, Y.-Y., Ho, T.-Y., and Jin, Y. Cloudleak: Large-scale deep learning models stealing through adversarial examples. In *NDSS*, 2020.

Yu, S., Sun, F., Guo, J., Zhang, R., and Cheng, X. Legonet: A fast and exact unlearning architecture. *arXiv preprint arXiv:2210.16023*, 2022b.

Zhang, B., Cai, T., Lu, Z., He, D., and Wang, L. Towards certifying l-infinity robustness using neural networks with l-inf-dist neurons. In *International Conference on Machine Learning*, pp. 12368–12379. PMLR, 2021.

Zhang, B., Jiang, D., He, D., and Wang, L. Rethinking lipschitz neural networks and certified robustness: A boolean function perspective. *Advances in Neural Information Processing Systems*, 35:19398–19413, 2022a.

Zhang, H., Chen, H., Song, Z., Boning, D. S., Dhillon, I. S., and Hsieh, C.-J. The limitations of adversarial training and the blind-spot attack. In *International Conference on Learning Representations*, 2019.

Zhang, Z., Zhou, Y., Zhao, X., Che, T., and Lyu, L. Prompt certified machine unlearning with randomized gradient smoothing and quantization. *Advances in Neural Information Processing Systems*, 35:13433–13455, 2022b.

Zhao, C., Qian, W., Ying, Z., and Huai, M. Static and sequential malicious attacks in the context of selective forgetting. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Zhou, W., Hou, X., Chen, Y., Tang, M., Huang, X., Gan, X., and Yang, Y. Transferable adversarial perturbations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.

Zhu, Y., Miao, C., Hajiaghajani, F., Huai, M., Su, L., and Qiao, C. Adversarial attacks against lidar semantic segmentation in autonomous driving. In *Proceedings of the 19th ACM conference on embedded networked sensor systems*, pp. 329–342, 2021.