Amplifying Membership Exposure via Data Poisoning

Anonymous Author(s)
Affiliation
Address
email

Abstract

As in-the-wild data are increasingly involved in the training stage, machine learning applications become more susceptible to data poisoning attacks. Such attacks typically lead to test-time accuracy degradation or controlled misprediction. In this paper, we investigate the third type of exploitation of data poisoning - increasing the risks of privacy leakage of benign training samples. To this end, we demonstrate a set of data poisoning attacks to amplify the membership exposure of the targeted class. We first propose a generic *dirty-label* attack for supervised classification algorithms. We then propose an optimization-based *clean-label* attack in the transfer learning scenario, whereby the poisoning samples are correctly labeled and look "natural" to evade human moderation. We extensively evaluate our attacks on computer vision benchmarks. Our results show that the proposed attacks can substantially increase the membership inference precision with minimum overall test-time model performance degradation.

1 Introduction

Training data are the most critical ingredients of machine learning, which are sometimes regarded as a new type of fuel in the era of artificial intelligence [21]. Over the last decade, training data collection and preservation have received growing concerns [20, 6], among which two focal agendas arise.

One is data corruption caused by *data poisoning* attacks, which pose serious threats to training data **integrity**. Data poisoning attacks exploit the common practice to (usually automatically) collect training data in the wild, e.g., from the Internet. This practice opens doors for attackers to inject malicious data at the training time to manipulate model behaviors at the test time. Consequently, models trained with poisoned training data suffer from either accuracy degradation [4, 13], targeted misclassification [25, 44], or backdoor implantation [23, 41].

The other is data leakage caused by *privacy inference* attacks, which mainly violate training data **confidentiality**. Ideally, a machine learning model learns generalizable knowledge of the training data, rather than details of specific data points. However, previous studies show that conventional learning algorithms can unintentionally remember sensitive information. Such information can later be inferred by various privacy inference attacks, such as membership inference attacks [29, 24].

Essentially, data poisoning is a training time attack while privacy inference is performed at the testing time. Despite intensive research efforts, most existing works study them separately. It remains unclear whether these two attacks can be integrated to cause more severe damage to machine learning models.

In this paper, we aim to advance the research frontier on the connection between data integrity and confidentiality. To this end, we propose a set of poisoning attacks to increase the precision of membership inference attacks, which is widely adopted as a standard tool to measure privacy leakage

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

in statistical analysis [31, 10]. In particular, we first show a simple but effective dirty-label poisoning attack that is generic for supervised classification applications. We then demonstrate a clean-label poisoning attack applicable to the transfer learning scenario. It has two significant advantages: (1) no requirements on the labeling process and (2) natural appearances to human moderation. Consequently, our proposed clean-label poisoning attack challenges the case where automatically crawling data from the Internet has become a common practice.

Our contributions are as follows: (1) We reveal an underexplored data poisoning attack threatening the training data privacy. In particular, we demonstrate how to amplify membership inference exposure of a specific class with only slight impacts on the model performance. (2) In generic supervised learning settings, we introduce a naïve dirty-label poisoning attack by modifying the labels of the poisoning samples. (3) In the transfer learning setting, we propose an optimization-based clean-label poisoning attack, wherein for each poisoning sample, we make imperceptible modifications without changing its label. It presents a more practical attack example since it does not require control over the labeling process. It is also a more stealthy attack as the contents of poisoning instances "look similar" to the natural ones. (4) We conduct extensive empirical studies to investigate the impacts of various factors, including datasets, architectures, poisoning budgets, and learning setups[1]. Our results show that our proposed poisoning attacks are able to increase the membership exposure obviously with just limited influences on the victim model's performance on testing samples.

2 Background

In this work, we focus on deep supervised classification in the computer vision field. That is, given a training dataset $\mathcal{D}_{\text{train}}$ composed of input-label pairs (x, y), a deep classifier $f(x;\theta)$ with model parameters θ, and a learning objective $\mathcal{L}\left(f(x;\theta), y\right)$, the learning process aims to find a set of model parameters θ^* to minimize the learning objective on $\mathcal{D}_{\text{train}}$, i.e., $\theta^* = \arg\min_{\theta} \sum_{(x,y)\in\mathcal{D}_{\text{train}}} \mathcal{L}\left(f(x;\theta), y\right)$.

2.1 Data Poisoning

In a poisoning attack, an attacker crafts and injects a set of malicious training samples $\mathcal{D}_{\text{poison}}$ (i.e., *poisoning dataset*) into the benign dataset $\mathcal{D}_{\text{clean}}$ (i.e., *clean dataset*). In the training stage, the model holder executes the machine learning algorithm on the full *poisoned dataset* $\mathcal{D}_{\text{train}} = \mathcal{D}_{\text{clean}} \cup \mathcal{D}_{\text{poison}}$ to obtain the trained model $f(x;\theta^*)$. In the inference stage, $f(x;\theta^*)$ tends to show unexpected behaviors on targeted inputs $(x, y) \in \mathcal{D}_{\text{target}}$. Most existing literature poses the poisoning attack as a bi-level optimization problem [5, 11]:

$$\mathcal{D}_{\text{poison}} = \arg\min_{\mathcal{D}} \sum_{(x,y)\in\mathcal{D}_{\text{target}}} \mathcal{A}\left(f(x;\theta^*)\right), \text{s.t. } \theta^* = \arg\min_{\theta} \sum_{(x,y)\in\mathcal{D}_{\text{clean}}\cup\mathcal{D}_{\text{poison}}} \mathcal{L}\left(f(x;\theta), y\right) \quad (1)$$

where $\mathcal{A}$ is the adversarial objective of the poisoning attack. Typically, there are three adversarial objectives: (1) in the accuracy degradation case, $\mathcal{D}_{\text{target}}$ contains all testing samples and $\mathcal{A} = -\mathcal{L}(f(x;\theta^*), y)$; (2) in the targeted misclassification case, $\mathcal{D}_{\text{target}}$ contains samples expected to be misclassified into the target class t and $\mathcal{A} = \mathcal{L}(f(x;\theta^*), t)$; (3) in the backdoor implantation case, $\mathcal{D}_{\text{target}}$ involves samples with a trigger δ, where inputs embedded with the trigger are to be classified into the target class t and $\mathcal{A} = \mathcal{L}(f(x \oplus \delta;\theta^*), t)$.

Based on the attacker's capability, poisoning attacks can be further grouped into two categories:

- **Dirty-label poisoning.** Most classical poisoning attacks require the control of the labeling process, where the attacker is allowed to modify the labels in $\mathcal{D}_{\text{poison}}$ [13, 39, 33]. Such attack is called *dirty-label poisoning*. Despite promising attack performance, dirty-label poisoning becomes impractical in supervised machine learning scenarios for two main reasons. First, in reality, it is a common practice that only unlabeled data (e.g., images and videos) are scraped by crawlers and then labeled by human moderators. Second, machine

[1]For reproducibility, we will release the source code after acceptance.

learning developers usually utilize anomaly detection algorithms to filter out wrongly labeled samples. Therefore, there exists a minimal chance that the data samples with modified labels are preserved in the dataset.

- **Clean-label poisoning.** To overcome the shortcomings of dirty-label poisoning, recent studies propose *clean-label poisoning* attacks [25, 43, 44]. A clean-label poisoning attack satisfies two properties. First, the labels of poisoning samples remain unchanged in the poisoning process. Second, to be inconspicuous, each poisoning sample x_p visually resembles a natural sample x_n, which is always constrained by a L_p-norm distance $\|x_p - x_n\|_p < \epsilon$.

Our setup. In this work, we investigate data poisoning under a newly identified and underexplored adversarial objective: *privacy leakage amplification*. In particular, we focus on one of the most representative privacy inference attacks: membership inference attack [29, 24]. We start from the dirty-label poisoning setup, which establishes a baseline and cornerstone for more advanced attacks. Then, we investigate attack methods in the clean-label poisoning setup to improve practicability. Note that there are concurrent works that propose to use poisoning attacks to cause privacy leakage [18, 34].

2.2 Membership Inference Attack

In a membership inference attack, an attacker aims to infer whether a specific sample (x, y) belongs to the training dataset $\mathcal{D}_{\text{train}}$ at the test time [29, 24, 16]. Unintended membership exposure causes catastrophic privacy loss for individuals. For example, in the real world, a data sample x can be a clinical record or an individual. Membership inference attacks enable the attackers to infer whether this clinical record or individual has been used to train a model associated with a certain disease. As such, these attacks are widely adopted as basic metrics to quantify the privacy exposure in statistical data analysis algorithms [8, 15]. Henceforth, we use membership inference attacks to demonstrate how to exploit data poisoning to amplify privacy leakage in this paper.

Based on the attacker's capability, membership inference attacks can be grouped into two categories:

- **Black-box membership inference.** In this case, the attacker distinguishes members and non-members only using model outputs [29, 28, 12]. This case is generic to most machine learning contexts. There are two attack strategies in general. The first strategy is *model-based* [29], where the attacker builds multiple shadow models to mimic the victim model, then utilizes them to construct a dataset with member/non-member labels, and finally trains a binary classifier to predict member/non-member. The second strategy is *metric-based* [24, 30], where the attacker compares a designed metric M_{mem}, such as correctness or entropy, with a predefined threshold τ to infer if a sample belongs to the training dataset.
- **White-box membership inference.** In this case, model parameters θ^* or even intermediate training information such as gradients $\frac{\partial \mathcal{L}}{\partial \theta}$ are observable by the attacker [19, 22, 14]. Such capability provides additional information supporting inference attacks, which is usually achievable in collaborative learning settings. Most white-box attacks are model-based, as the attacker needs to access the internals of deep models to extract model-specific features.

Our setup. Our evaluation is carried out in the metric-based black-box membership inference setting. We adopt the following metric proposed by Song et al. [31] to help measure the membership exposure[2]:

$$M_{\text{mem}} = -\left(1 - f(x)_y\right) \log\left(f(x)_y\right) - \sum_{i \neq y} \log\left(1 - f(x)_i\right) f(x)_i \tag{2}$$

where $f(x)_j$ refers to the confidence value of label j. Equation 2 simultaneously considers the correctness and entropy metric. In the training stage, the learning algorithm continuously fits the training samples, by decreasing the entropy loss (i.e., the learning objective) and increasing the confidence score of the correct label. As a result, a member is likely to produce a higher M_{mem} than a non-member. In our evaluation, we adopt the AUC score to measure membership exposure when the threshold τ varies.

[2]For simplicity, we use the notation $f(x)$ to represent the trained model $f(x; \theta^*)$.

2.3 Threat Model

Before diving into our attack design, we first elaborate on the threat model considered in this paper.

Attack Goals. The first goal of the attacker is *increasing the chance of leaking the membership of training samples within a targeted class*. At the same time, the attacker attempts to make the poisoning attack as stealthy as possible. As such, the second goal of the attacker is *generating poisoning samples that have limited impact on the model performance for untargeted classes and are indistinguishable to natural samples*.

Attacker Capabilities. We assume the attacker has the basic capabilities set up by existing data poisoning and membership inference attack games. The attacker can craft poisoning samples and inject them into the victim's clean dataset $\mathcal{D}_{\text{clean}}$. However, there exists a *poisoning budget* b_{poison} to limit the amount of training samples ($|\mathcal{D}_{\text{poison}}| \leq b_{\text{poison}} \ll |\mathcal{D}_{\text{clean}}|$). We assume the attacker owns a *shadow dataset* $\mathcal{D}_{\text{shadow}}$ to craft poisoning samples, which contains natural samples from the same distribution with $\mathcal{D}_{\text{clean}}$. Moreover, the attacker cannot modify poisoning data labels in the clean-label poisoning setting. But we assume the attacker knows the feature extractor (i.e., an encoder) used by the victim model. This assumption is practical since developing models with public pretrained feature extractors is a common practice in existing clean-label poisoning literature [25, 43].

Remark. In our setup, the clean dataset $\mathcal{D}_{\text{clean}}$ is unknown by the attacker, which is different from most existing data poisoning games [4, 13, 18].

3 Attack Methodology

Although Equation 1 establishes a generic framework for poisoning attack, it is unsuitable for our case. First, classic solutions for the optimization problem posed by Equation 1 require differentiation to the inner loss minimizer, which is intractable to models with high complexity. Second, although meta-learning algorithms have been proposed to solve the bi-level optimization problem in the deep learning setting [11, 42, 32], they lead to huge computing costs, limiting the practicability of the attack. Third, in our threat model, $\mathcal{D}_{\text{clean}}$ is unobservable, making Equation 1 no longer applicable to our attack. Facing these obstacles, we explore heuristic strategies to achieve the attack goals.

3.1 Dirty-label Poisoning Attack: A Naïve Approach

We start our investigation in the dirty-label poisoning setting. This case has more limitations in the real world as discussed in Section 2.1. But it relaxes the attack constraints and helps to verify the feasibility of our attack strategy.

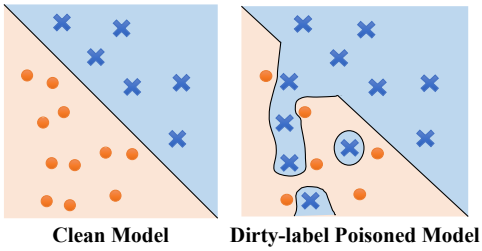


Figure 1: Conceptual illustration of decision the boundary of models trained without/with our dirty-label poisoning attack. To achieve a low training accuracy, the model has to learn a more complicated decision boundary, making the learning process more susceptible to overfitting.

The key to amplifying membership exposure is to cause overfitting in the targeted class. We propose the following attack strategy:

Our dirty-label poisoning attack. Given the label of the target group t , the shadow dataset $\mathcal{D}_{\text{shadow}}$, and a poisoning budget b_{poison}, the poisoning dataset is constructed by the following steps: (1) select all samples $(x, t) \in \mathcal{D}_{\text{shadow}}$ with label t; (2) for each sample (x, t), randomly change the label to another class $i \neq t$; (3) preserve at most b_{poison} samples (x, i) to compose $\mathcal{D}_{\text{poison}}$.

Intuitively, only learning general concepts/features is insufficient to discriminate the clean samples of class t from poisoning samples, since in fact, they share similar features. To minimize the training loss, the victim model has to learn more specific features of each sample. As conceptually illustrated by Figure 1, the generalization performance on class t tend to degrade. Figure 2 gives an example to

show how our poisoning attack improves the membership exposure of the *airplane* category for an InceptionV3-based CIFAR-10 classifier. In this example, we assume the numbers of members and non-members are both 10,000 and b_{poison}=1,000.

Remark. We do not pose extra assumptions to the training process in this case. Our dirty-label poisoning attack is generic to all supervised learning scenarios, including end-to-end learning and transfer learning. Similar to the concurrent work [34], we also use the classical *label flipping* strategy [4] to carry out the dirty-label attack.

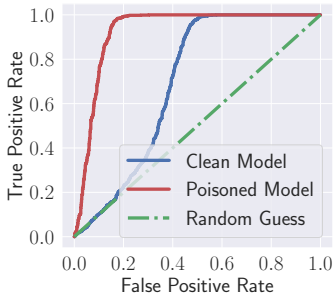


Figure 2: Membership inference against the *airplane* class of CIFAR-10 dataset. The poisoning budget is set as $b_{\text{poison}} = 10\% \times |\mathcal{D}_{\text{clean}}|$. The AUC score, i.e., the area under the ROC curve, is increased from 0.6917 to 0.9255, indicating that our dirty-label poisoning attack is effective.

3.2 Clean-label Poisoning Attack: A Stealthy Approach

In this part, we aim to improve our poisoning attack to amplify membership exposure while attaining stealthiness. Following the basic setup in [25], we consider the transfer learning setting. In transfer learning, we assume the victim model is composed of two parts: a pretrained feature extractor $g(\cdot)$ to extract high-level features from the input x and a newly-trained classifier $c(\cdot)$ to predict labels.

The key to our clean-label poisoning is to exploit one vulnerability of deep feature extractors: a slight change in the input space may cause a significant change in the feature space. We propose the following attack strategy.

Our clean-label poisoning attack. Given the label of the target class t , the shadow dataset $\mathcal{D}_{\text{shadow}}$, a poisoning budget b_{poison}, and the adopted feature extractor g, the poisoning dataset is constructed by the following steps: (1) from $\mathcal{D}_{\text{shadow}}$, select a *base sample* (x_{base}, t) with label t and a sample (x, y) where $y \neq t$; (2) find a x^* that is close to x in the input space ($\|x^* - x\|_p \leq \epsilon$) and close to x_{base} in the feature space ($g(x^*) \approx g(x_{\text{base}})$), and insert (x^*, y) to $\mathcal{D}_{\text{poison}}$; (3) repeat step (1) and (2) until $|\mathcal{D}_{\text{poison}}| = b_{\text{poison}}$.

The intuition of our clean-label poisoning attack is straightforward: we actually mount a dirty-label poisoning attack in the feature space. That is the poisoning sample (x^*, y) is close to clean samples in class t in the feature space but labeled as y. We formalize step (2) as a constrained optimization problem:

$$x^* = \arg\min_{x'} \|g(x') - g(x_{\text{base}})\|_2, \text{s.t. } \|x' - x_{\text{base}}\|_\infty \leq \epsilon \text{ and } x' \in \mathcal{X} \tag{3}$$

where $\mathcal{X}$ refers to the input space. For instance, for a normalized RGB image input, a valid pixel value in one color channel should be within $[0, 1]$. For ease of optimization, we adopt L_2-norm to measure the distance in the feature space. Following the convention of adversarial attack work, we adopt L_∞ to measure the distance in the input space. In our case, the constraint can be expressed as $x' \in [x_{\text{min}}, x_{\text{max}}]$, where $x_{\text{min}} = \max(0, x_{\text{base}} - \epsilon)$, $x_{\text{max}} = \min(1, x_{\text{base}} + \epsilon)$.

Optimization. We adopt the variable substitution method introduced by [7] to solve Equation 3. We introduce a new variable w and let

$$\tanh(w) = \frac{2(x' - x_{\text{min}})}{x_{\text{max}} - x_{\text{min}}} - 1 \tag{4}$$

Based on the fact that $\tanh(w) \in [-1, 1]$, we can prove that the constraint $x' \in [x_{\text{min}}, x_{\text{max}}]$ holds in Equation 4. Plugging Equation 4 into Equation 3, we can obtain

$$\begin{aligned} w^* &= \arg\min_{w} \|g\left(0.5 * (\tanh(w) + 1)(x_{\text{max}} - x_{\text{min}}) + x_{\text{min}}\right) - g(x_{\text{base}})\|_2, \\ \text{and } \; x^* &= 0.5(\tanh(w^*) + 1)(x_{\text{max}} - x_{\text{min}}) + x_{\text{min}} \end{aligned} \tag{5}$$

which poses an unconstrained optimization problem that can be simply solved by mainstream gradient optimization methods. In our implementation, we first initialize variable w with x_{base} by Equation 4 and iterate the optimization process 1,000 times, using an Adam optimizer with the learning rate 0.01.

Figure 3 illustrates some examples generated by our proposed two poisoning attacks. We can see that the poisoning samples produced by our clean-label poisoning method look natural to human eyes, which are hard to detect by manual moderation.

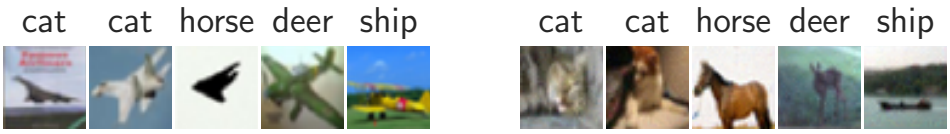


Figure 3: Examples of poisoning samples against the *airplane* category. **Left:** Dirty-label poisoning samples. **Right:** Clean-label poisoning samples (ϵ=16/255). We provide the label of each sample on the top. For the clean-label poisoning samples, they look like natural ones.

4 Evaluation

4.1 Experimental Setup

Datasets. In our experiment, we split each dataset into three portions: the clean training dataset $\mathcal{D}_{\text{clean}}$ containing the members, the test dataset $\mathcal{D}_{\text{test}}$ containing the non-members, and the shadow dataset $\mathcal{D}_{\text{shadow}}$ for generating poisoning samples. We follow the same setup with [29], where $|\mathcal{D}_{\text{clean}}| = |\mathcal{D}_{\text{test}}| = |\mathcal{D}_{\text{shadow}}|$, and each of them does not overlap with others. Additionally, we set the three datasets to be balanced for the simplicity of evaluation among each class. We adopt five datasets for our experiments, including (1) **MNIST** [1] that contains 60,000 handwritten digits from 0 to 9. (2) **CIFAR-10** [2] that contains 60,000 images from 10 classes. (3) **STL-10** [3] that contains 130,000 labeled images from 10 classes. (4) **CelebA** [17] that contains 202,599 face images annotated by 40 attributes. In our experiment, we train a two-class classifier is to predict the *Charming* attribute; (5) **PatchCamelyon** [37] that contains 327,680 images to predict presence of metastatic tissue. For MNIST, CIFAR-10, CelebA, and PatchCamelyon dataset, we set $|\mathcal{D}_{\text{clean}}|$=10,000, while for STL-10, we set $|\mathcal{D}_{\text{clean}}|$=4,000. Each sample is transformed into a $96 \times 96 \times 3$ RGB image.

Models. We use five pretrained models provided by Tensorflow(v2.5.2): Xception, ResNet18, MobileNetv2, InceptionV3, and VGG16. For each model, we remove the fully connected (FC) layers to build up the feature extractor $g(\cdot)$. Then, we add two FC layers on top of the feature extractor to form $c(\cdot)$: one layer with 128 hidden units using the Tanh activation function, followed by an output layer using the Softmax activation function. We fix the feature extractor and train the newly added FC layers with the Adam optimizer, with the learning rate of 10^{-5} and batch size of 100.

Poisoning Setup. For each learning task, we mount the poisoning attack against each class. We set the poisoning budget b_{poison} as $\frac{|\mathcal{D}_{\text{clean}}|}{\#\text{classes}}$. To make our poisoning attack stealthy to the victim, we assume $\mathcal{D}_{\text{poison}}$ is evenly distributed among all classes with each having approximately $\frac{b_{\text{poison}}}{\#\text{classes}}$ poisoned data samples. For instance, for a CIFAR-10 classifier, we set $b_{\text{poison}} = 1,000$, where only 100 samples for each class are in $\mathcal{D}_{\text{poison}}$. Note that this assumption means that there are $\frac{b_{\text{poison}}}{\#\text{classes}}$ clean samples in $\mathcal{D}_{\text{clean}}$ for the target class t. For the clean-label poisoning dataset, we set the perturbation constraint $\epsilon = 16/255$.

Equipment. Our experiments were conducted on a deep learning server, which equipped with an Intel(R) Xeon(R) Gold 6226R CPU @ 2.90GHz, 128GB RAM, and four NVIDIA GeForce RTX 3090 GPUs with 24GB of memory.

	Without Poisoning		Dirty-Label Poisoning		Clean-Label Poisoning	
	MI AUC	Test Acc.	MI AUC	Test Acc.	MI AUC	Test Acc.
MNIST						
Xception	.538±.022	.939	.697±.020	.918±.007	.627±.032	.924±.005
InceptionV3	.546±.029	.928	.791±.026	.902±.002	.674±.053	.910±.005
VGG16	.536±.021	.954	.740±.029	.934±.004	.604±.033	.943±.002
ResNet50	.525±.021	.967	.721±.034	.946±.009	.583±.022	.963±.002
MobileNetV2	.529±.019	.960	.759±.043	.937±.004	.626±.039	.949±.004
CIFAR-10						
Xception	.642±.057	.769	.893±.018	.735±.004	.868±.032	.738±.004
InceptionV3	.733±.058	.676	.935±.015	.648±.005	.827±.046	.663±.002
VGG16	.619±.049	.815	.899±.011	.779±.004	.869±.019	.783±.003
ResNet50	.597±.035	.849	.930±.015	.832±.003	.861±.042	.838±.002
MobileNetV2	.602±.050	.842	.916±.011	.813±.002	.836±.037	.820±.002
STL-10						
Xception	.578±.043	.857	.906±.021	.836±.004	.876±.027	.838±.004
InceptionV3	.696±.069	.758	.940±.020	.735±.006	.809±.065	.750±.002
VGG16	.596±.040	.875	.895±.017	.842±.004	.858±.029	.846±.005
ResNet50	.572±.031	.897	.931±.015	.875±.005	.860±.041	.883±.003
MobileNetV2	.558±.033	.935	.900±.019	.907±.004	.822±.042	.916±.004
CelebA						
Xception	.644±.011	.724	.752±.001	.693±.002	.710±.027	.689±.005
InceptionV3	.748±.019	.687	.849±.008	.655±.000	.759±.056	.661±.001
VGG16	.711±.019	.724	.747±.005	.684±.003	.713±.005	.680±.003
ResNet50	.571±.000	.744	.680±.007	.698±.007	.616±.049	.713±.000
MobileNetV2	.680±.009	.750	.823±.002	.704±.002	.742±.031	.696±.006
PatchCamelyon						
Xception	.564±.003	.847	.678±.020	.797±.002	.644±.007	.816±.001
InceptionV3	.617±.008	.832	.739±.029	.774±.008	.627±.038	.800±.001
VGG16	.538±.003	.862	.623±.016	.842±.004	.593±.004	.838±.001
ResNet50	.543±.005	.891	.701±.038	.820±.015	.611±.017	.862±.002
MobileNetV2	.565±.004	.890	.728±.029	.819±.000	.674±.001	.842±.007

Table 1: Membership inference (MI) AUC of the target class t and test-time accuracy on $\mathcal{D}_{\text{test}}$, for models not poisoned, poisoned by dirty-label attacks, and poisoned by clean-label attacks, respectively. We run the attack and evaluation over each class, and we report the average value with standard deviation for each metric.

4.2 Results

Table 1 reports the evaluation results on our proposed two poisoning attacks. In general, we can observe that our attacks obviously increase the AUC score of the membership inference against the target class t, with slight testing accuracy decay. For instance, for the clean-label poisoning attack against the STL-10 Xception-based classifier, the membership inference AUC is increased from 0.578 to 0.870 on average, while the testing accuracy is decreased from 0.861 to 0.837 on average. The results imply that our clean-poisoning attack is hard to detect: the poisoned samples resemble the clean samples and the performance of the poisoned models degrades subtly on the testing dataset.

Another observation is that the dirty-label poisoning attack causes more significant membership exposure than the clean-label poisoning attack. A potential explanation of this phenomenon is that our clean-label poisoning attack can be considered as an "approximate" version of the dirty-label poisoning attack in the feature space.

5 Ablation Study

5.1 Study 1: Impact of $\mathcal{D}_{\text{shadow}}$

In our prior study, we assume the attacker can obtain a shadow dataset $\mathcal{D}_{\text{shadow}}$ with the same distribution of the victim's clean training dataset $\mathcal{D}_{\text{clean}}$. However, the model developer may hide the training dataset information to fortify privacy and intellectual property (IP) protection. In this ablation section, we examine how the $\mathcal{D}_{\text{shadow}}$ affects the attack performance.

In our experiment, we assume the attacker aims to poison the *cat* and *aircraft* categories against the CIFAR-10 classifiers respectively. For each target class t, we select 1,000 samples from t in STL-10. In this ablation study, We start with the dirty-label poisoning attack to estimate the upper bound performance of our clean-label poisoning attack. Table 2 reports the attack performance. Our results show that the attack performance declines when the attacker's shadow data has a different distribution from the actual clean dataset.

	$t = 0$ (*airplane*)		$t = 3$ (*cat*)	
	Δ MI AUC	Δ Test Acc.	Δ MI AUC	Δ Test Acc.
Xception	.056	-.004	.074	-.010
InceptionV3	.023	-.017	.042	-.006
VGG16	.006	-.007	.035	-.006
ResNet50	.015	-.008	.041	-.010
MobileNetV2	.044	-.006	.007	-.006

Table 2: Dirty-poisoning attack results when $\mathcal{D}_{\text{clean}}$ comes from CIFAR-10 and $\mathcal{D}_{\text{shadow}}$ contains 1,000 instances from the target class t of STL-10. We report the membership inference (MI) AUC difference and testing accuracy between the poisoned model and the clean model.

5.2 Study 2: Impact of b_{poison}

We then study how the size of poisoning samples affects the attack performance. Figure 4 depicts the membership exposure under dirty-label poisoning attacks with different poisoning budget b_{poison}. Unsurprisingly, as we gradually decrease the poisoning budget, the membership inference AUC score decreases as well. Yet, it is still possible to achieve a significant membership exposure increase even with a small poison budget. Take ResNet50 for instance, when b_{poison} is 100, i.e., the amount of poisoning samples is only 1% of that for the clean samples, we can improve the membership inference AUC score from 0.5971 to 0.7076 on average.

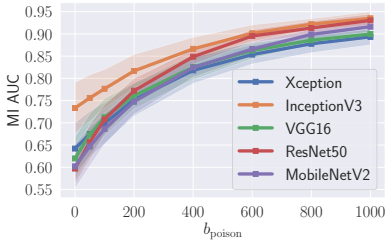


Figure 4: Membership exposure under different b_{poison}. We plot the average MI AUC on the CIFAR-10 dataset.

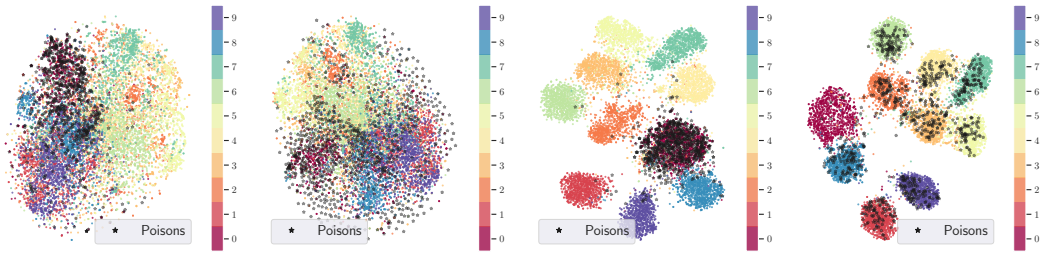
5.3 Impact of Fine-tuning

In the transfer learning scenario, there are also chances that model developers fine-tune parameters of feature extractors [38, 9, 27]. The fine-tuning operation is supposed to make the transfer learning model better fit the new dataset. Inevitably, it may affect how the feature extractor extracts latent features from the inputs. As a result, our poisoning attack may get affected. In this part, we explore how the fine-tuning process impacts our attack performance.

In our experiment, we train CIFAR-10 classifiers with fine-tuning on the poisoned dataset, with the learning rate set as 1e-5. We report the attack performance in Table 3. It can be seen that, for our dirty-label poisoning attack, we achieve more significant membership exposure amplifications. While for our clean-label poisoning attack, the membership exposure changes are nearly negligible.

	Without Poisoning		Dirty-Label Poisoning		Clean-Label Poisoning	
	MI AUC	Test Acc.	MI AUC	Test Acc.	MI AUC	Test Acc.
Xception	.573±.030	.899	.923±.007	.878±.002	.581±.030	.898±.001
InceptionV3	.619±.053	.863	.956±.008	.845±.002	.618±.054	.863±.001
VGG16	.622±.052	.875	.904±.010	.858±.002	.622±.060	.874±.002
ResNet50	.584±.030	.903	.942±.009	.884±.002	.592±.032	.901±.001
MobileNetV2	.553±.026	.906	.820±.020	.896±.002	.556±.024	.909±.001

Table 3: Dirty-label and clean-label attack results on fine-tuned CIFAR-10 classifier ($b_{\text{poison}} = 1,000$). We report the membership inference (MI) AUC difference and testing accuracy between the poisoned model and the clean model.



(a) (b) (c) (d)

Figure 5: Visualization of latent features extracted by different poisoned CIFAR-10 classifiers from the InceptionV3 feature extractor. The targeted class is 0 (*airplane*). (a) Dirty-label attack against a fixed feature extractor. (b) Clean-label attack against a fixed feature extractor. (c) Dirty-label attack against a fine-tuned model. (d) Clean-label attack against a fine-tuned model. The colored points are the clean training samples, while the dark star marks are the poisoning sample.

To understand this phenomenon, we investigate how the feature extractor behaves in different cases. We visualize the latent features of clean samples and poisoning samples by t-SNE technique [36] in Figure 5. We can observe that, for the dirty-label attack in the fine-tuning setting, the poisoning samples distribute closely to the targeted clean samples in the feature space (as shown in Figure 5c. Besides, for the clean-label attack in the fine-tuning setting, the poisoning samples distribute among each class in the feature space (as shown in Figure 5d). It indicates that the more poisoning samples in the target class, the more significant membership exposure results we can get. By comparing Figure 5a and Figure 5b, we can also partially illustrate that dirty-label poisoning tends to achieve better attack performance than the clean-label poisoning. This trend is actually consistent with the intuition of our attack. In fact, this is the case called "adversarial training" by some adversarial example defense literature [35, 26, 40]. During adversarial training, model parameters are gradually tuned to eliminate the effects of input perturbations.

6 Conclusion

In this paper, we demonstrate the feasibility of exploiting data poisoning to amplify the membership exposure of the training dataset. We present one significantly increase the precision of the membership inference attack on classes with limited negative influences on the model's test-time performance. We also conduct extensive evaluations to study how different factors may affect attack performance.

Our findings uncover a new challenge to modern machine learning ecosystems. Data from the open world may not only affect the model performance but also raise real threats to private training data. Even worse, we show it is possible to mount clean-label attacks to evade human moderation. One primary limitation of our clean-label attacks is that we assume the attacker has knowledge of the feature extractor. How to create clean-label poisons in a black-box manner and how to defend such privacy-violating poisoning attacks remain our future work.

References

[1] http://yann.lecun.com/exdb/mnist/.

[2] https://www.cs.toronto.edu/~kriz/cifar.html.

[3] https://cs.stanford.edu/%7Eacoates/stl10/.

[4] Battista Biggio, Blaine Nelson, and Pavel Laskov. Poisoning Attacks against Support Vector Machines. In *International Conference on Machine Learning (ICML)*. icml.cc / Omnipress, 2012.

[5] Aleksandar Bojchevski and Stephan Günnemann. Adversarial Attacks on Node Embeddings via Graph Poisoning. In *International Conference on Machine Learning (ICML)*, pages 695–704. PMLR, 2019.

[6] Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks. In *USENIX Security Symposium (USENIX Security)*, pages 267–284. USENIX, 2019.

[7] Nicholas Carlini and David Wagner. Towards Evaluating the Robustness of Neural Networks. In *IEEE Symposium on Security and Privacy (S&P)*, pages 39–57. IEEE, 2017.

[8] Cynthia Dwork and Aaron Roth. *The Algorithmic Foundations of Differential Privacy*. Now Publishers Inc., 2014.

[9] Yunhui Guo, Honghui Shi, Abhishek Kumar, Kristen Grauman, Tajana Rosing, and Rogério Schmidt Feris. Spottune: Transfer learning through adaptive fine-tuning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4805–4814. IEEE, 2019.

[10] Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. LOGAN: Evaluating Privacy Leakage of Generative Models Using Generative Adversarial Networks. *Symposium on Privacy Enhancing Technologies Symposium*, 2019.

[11] W. Ronny Huang, Jonas Geiping, Liam Fowl, Gavin Taylor, and Tom Goldstein. MetaPoison: Practical General-purpose Clean-label Data Poisoning. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 12080–12091. NeurIPS, 2020.

[12] Bo Hui, Yuchen Yang, Haolin Yuan, Philippe Burlina, Neil Zhenqiang Gong, and Yinzhi Cao. Practical Blind Membership Inference Attack via Differential Comparisons. In *Network and Distributed System Security Symposium (NDSS)*. Internet Society, 2021.

[13] Matthew Jagielski, Alina Oprea, Battista Biggio, Chang Liu, Cristina Nita-Rotaru, and Bo Li. Manipulating Machine Learning: Poisoning Attacks and Countermeasures for Regression Learning. In *IEEE Symposium on Security and Privacy (S&P)*, pages 19–35. IEEE, 2018.

[14] Klas Leino and Matt Fredrikson. Stolen Memories: Leveraging Model Memorization for Calibrated White-Box Membership Inference. In *USENIX Security Symposium (USENIX Security)*, pages 1605–1622. USENIX, 2020.

[15] Ninghui Li, Wahbeh H. Qardaji, Dong Su, Yi Wu, and Weining Yang. Membership Privacy: A Unifying Framework for Privacy Definitions. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 889–900. ACM, 2013.

[16] Zheng Li and Yang Zhang. Membership Leakage in Label-Only Exposures. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2021.

[17] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep Learning Face Attributes in the Wild. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738. IEEE, 2015.

[18] Saeed Mahloujifar, Esha Ghosh, and Melissa Chase. Property Inference from Poisoning . In *IEEE Symposium on Security and Privacy (S&P)*, pages 1569–1569. IEEE, 2022.

[19] Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning. In *IEEE Symposium on Security and Privacy (S&P)*, pages 1021–1035. IEEE, 2019.

[20] Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael Wellman. SoK: Towards the Science of Security and Privacy in Machine Learning. In *IEEE European Symposium on Security and Privacy (Euro S&P)*, pages 399–414. IEEE, 2018.

[21] Yuji Roh, Geon Heo, and Steven Euijong Whang. A survey on data collection for machine learning: A big data - ai integration perspective. *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[22] Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, Yann Ollivier, and Hervé Jégou. White-box vs Black-box: Bayes Optimal Strategies for Membership Inference. In *International Conference on Machine Learning (ICML)*, pages 5558–5567. PMLR, 2019.

[23] Ahmed Salem, Rui Wen, Michael Backes, Shiqing Ma, and Yang Zhang. Dynamic Backdoor Attacks Against Machine Learning Models. *CoRR abs/2003.03675*, 2020.

[24] Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models. In *Network and Distributed System Security Symposium (NDSS)*. Internet Society, 2019.

[25] Ali Shafahi, W Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 6103–6113. NeurIPS, 2018.

[26] Ali Shafahi, Mahyar Najibi, Amin Ghiasi, Zheng Xu, John P. Dickerson, Christoph Studer, Larry S. Davis, Gavin Taylor, and Tom Goldstein. Adversarial Training for Free! In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, pages 3353–3364. NeurIPS, 2019.

[27] Ali Shafahi, Parsa Saadatpanah, Chen Zhu, Amin Ghiasi, Christoph Studer, David W. Jacobs, and Tom Goldstein. Adversarially robust transfer learning. In *International Conference on Learning Representations (ICLR)*, 2020.

[28] Reza Shokri, Martin Strobel, and Yair Zick. Exploiting Transparency Measures for Membership Inference: a Cautionary Tale. In *The AAAI Workshop on Privacy-Preserving Artificial Intelligence (PPAI)*. AAAI, 2020.

[29] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership Inference Attacks Against Machine Learning Models. In *IEEE Symposium on Security and Privacy (S&P)*, pages 3–18. IEEE, 2017.

[30] Liwei Song and Prateek Mittal. Systematic Evaluation of Privacy Risks of Machine Learning Models. In *USENIX Security Symposium (USENIX Security)*, pages 2615–2632. USENIX, 2021.

[31] Liwei Song, Reza Shokri, and Prateek Mittal. Privacy Risks of Securing Machine Learning Models against Adversarial Examples. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 241–257. ACM, 2019.

[32] Zheng Tianhang and Li Baochun. First-Order Efficient General-Purpose Clean-Label Data Poisoning . In *IEEE Conference on Computer Communications (INFOCOM)*. IEEE, 2021.

[33] Vale Tolpegin, Stacey Truex, Mehmet Emre Gursoy, and Ling Liu. Data Poisoning Attacks Against Federated Learning Systems. In *European Symposium on Research in Computer Security (ESORICS)*, pages 480–501. Springer, 2020.

[34] Florian Tramè, Reza Shokri, Ayrton San Joaquin, Hoang Le, Matthew Jagielski, Sanghyun Hong, and Nicholas Carlini. Truth Serum: Poisoning Machine Learning Models to Reveal Their Secrets. *CoRR abs/2204.00032*, 2022.

[35] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble Adversarial Training: Attacks and Defenses. In *International Conference on Learning Representations (ICLR)*, 2017.

[36] Laurens van der Maaten and Geoffrey Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 2008.

[37] Bastiaan S Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation Equivariant CNNs for Digital Pathology. *CoRR abs/1806.03962*, 2018.

[38] Bolun Wang, Yuanshun Yao, Bimal Viswanath, Haitao Zheng, and Ben Y. Zhao. With Great Training Comes Great Vulnerability: Practical Attacks against Transfer Learning. In *USENIX Security Symposium (USENIX Security)*, pages 1281–1297. USENIX, 2018.

[39] Hongyi Wang, Kartik Sreenivasan, Shashank Rajput, Harit Vishwakarma, Saurabh Agarwal, Jy yong Sohn, Kangwook Lee, and Dimitris Papailiopoulos. Attack of the Tails: Yes, You Really Can Backdoor Federated Learning. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*. NeurIPS, 2020.

[40] Eric Wong, Leslie Rice, and J. Zico Kolter. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations (ICLR)*, 2020.

[41] Zhaohan Xi, Ren Pang, Shouling Ji, and Ting Wang. Graph Backdoor. In *USENIX Security Symposium (USENIX Security)*. USENIX, 2021.

[42] Hengtong Zhang, Jing Gao, and Lu Su. Data Poisoning Attacks Against Outcome Interpretations of Predictive Models. In *ACM Conference on Knowledge Discovery and Data Mining (KDD)*, page 2165–2173. ACM, 2021.

[43] Shihao Zhao, Xingjun Ma, Xiang Zheng, James Bailey, Jingjing Chen, and Yu-Gang Jiang. Clean-Label Backdoor Attacks on Video Recognition Models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14443–144528. IEEE, 2020.

[44] Chen Zhu, W. Ronny Huang, Hengduo Li, Gavin Taylor, Christoph Studer, and Tom Goldstein. Transferable Clean-Label Poisoning Attacks on Deep Neural Nets. In *International Conference on Machine Learning (ICML)*, pages 7614–7623. PMLR, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 6.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 6.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...

(a) Did you state the full set of assumptions of all theoretical results? [N/A]

(b) Did you include complete proofs of all theoretical results? [N/A]

3. If you ran experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We will submit the code, link to data we used, and instructions in the supplemental material.

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 4.1.

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] We reset the random seed (54321) at the beginning of building clean models and poisoned models.

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 4.1.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes] We cite the data and models we used in Section 4.1.

(b) Did you mention the license of the assets? [Yes]

(c) Did you include any new assets either in the supplemental material or as a URL? [No] We did not use any new assets.

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]