

---

# Just How Toxic is Data Poisoning? A Unified Benchmark for Backdoor and Data Poisoning Attacks

---

Avi Schwarzschild<sup>\*1</sup> Micah Goldblum<sup>\*2</sup> Arjun Gupta<sup>3</sup> John P. Dickerson<sup>2</sup> Tom Goldstein<sup>2</sup>

## Abstract

Data poisoning and backdoor attacks manipulate training data in order to cause models to fail during inference. A recent survey of industry practitioners found that data poisoning is the number one concern among threats ranging from model stealing to adversarial attacks. However, it remains unclear exactly how dangerous poisoning methods are and which ones are more effective considering that these methods, even ones with identical objectives, have not been tested in consistent or realistic settings. We observe that data poisoning and backdoor attacks are highly sensitive to variations in the testing setup. Moreover, we find that existing methods may not generalize to realistic settings. While these existing works serve as valuable prototypes for data poisoning, we apply rigorous tests to determine the extent to which we should fear them. In order to promote fair comparison in future work, we develop standardized benchmarks for data poisoning and backdoor attacks.

## 1. Introduction

*Data poisoning* is a security threat to machine learning systems in which an attacker controls the behavior of a system by manipulating its training data. This class of threats is particularly germane to deep learning systems because they require large amounts of data to train and are therefore often trained (or pre-trained) on large datasets scraped from the web. For example, the Open Images and the Amazon Products datasets contain approximately 9 million and 233 million samples, respectively, that are scraped from a wide range of potentially insecure, and in many cases unknown, sources (Kuznetsova et al., 2020; Ni et al., 2019).

---

<sup>1</sup>Department of Mathematics, <sup>2</sup>Department of Computer Science, and <sup>3</sup>Department of Robotics, University of Maryland, College Park, MD, USA. Correspondence to: Avi Schwarzschild <avi1@umd.edu>.

At this scale, it is often infeasible to properly vet content. Furthermore, many practitioners create datasets by harvesting system inputs (e.g., emails received, files uploaded) or scraping user-created content (e.g., profiles, text messages, advertisements) without any mechanisms to bar malicious actors from contributing data. The dependence of industrial AI systems on datasets that are not manually inspected has led to fear that corrupted training data could produce faulty models (Jiang et al., 2017). In fact, a recent survey of 28 industry organizations found that these companies are significantly more afraid of data poisoning than other threats from adversarial machine learning (Kumar et al., 2020b).

Poisoning attacks can be put into two broad categories. *Backdoor data poisoning* causes a model to misclassify test-time samples that contain a *trigger* – a visual feature in images or a particular character sequence in the natural language setting (Chen et al., 2017; Dai et al., 2019; Saha et al., 2019; Turner et al., 2018). For example, one might tamper with training images so that a vision system fails to identify any person wearing a shirt with the trigger symbol printed on it. In this threat model, the attacker modifies data at both train time (by placing poisons) and at inference time (by inserting the trigger). *Triggerless* poisoning attacks, on the other hand, do not require modification at inference time (Biggio et al., 2012; Huang et al., 2020; Muñoz-González et al., 2017; Shafahi et al., 2018; Zhu et al., 2019; Aghakhani et al., 2020b; Geiping et al., 2020). A variety of innovative backdoor and triggerless poisoning attacks – and defenses – have emerged in recent years, but inconsistent and perfunctory experimentation has rendered performance evaluations and comparisons misleading.

In this paper, we develop a framework for benchmarking and evaluating a wide range of poison attacks on image classifiers. Specifically, we provide a way to compare attack strategies and shed light on the differences between them.

Our goal is to address the following weaknesses in the current literature. First, we observe that the reported success of poisoning attacks in the literature is often dependent on specific (and sometimes unrealistic) choices of network architecture and training protocol, making it difficult to assess the viability of attacks in real-world scenarios. Second, we find that the percentage of training data that an attacker can

modify, the standard budget measure in the poisoning literature, is not a useful metric for comparisons. The flaw in this metric invalidates comparisons because even with a fixed percentage of the dataset poisoned, the success rate of an attack can still be strongly dependent on the dataset size, which is not standardized across experiments to date. Third, we find that some attacks that claim to be “clean label,” such that poisoned data still appears natural and properly labeled upon human inspection, are not.

Our proposed benchmarks measure the effectiveness of attacks in standardized scenarios using modern network architectures. We benchmark from-scratch training scenarios and also white-box and black-box transfer learning settings. Also, we constrain poisoned images to be *clean* in the sense of small perturbations. Furthermore, our benchmarks are publicly available as a proving ground for existing and future data poisoning attacks.

The data poisoning literature contains attacks in a variety of settings including image classification, facial recognition, and text classification (Shafahi et al., 2018; Chen et al., 2017; Dai et al., 2019). Attacks on the fairness of models, speech recognition, and recommendation engines have also been developed (Solans et al., 2020; Aghakhani et al., 2020a; Li et al., 2016; Fang et al., 2018; Hu et al., 2019; Fang et al., 2020). In addition to a variety of applications, the threat models range from attackers having access only to data all the way to the attacker controlling the entire training process (Gu et al., 2017; Yao et al., 2019; Salem et al., 2020a;b).

While we acknowledge the merits of studying poisoning in a range of modalities, our benchmark focuses on attacks on image classifiers that only modify data since this is by far the most common setting in the existing literature, and even among these attacks, there has not yet been a standard comparison metric. Specifically, we focus on attacks with a common goal and the sensitivities to experimental setup that we explore are not deviations from this goal.

## 2. A Synopsis of Triggerless and Backdoor Data Poisoning

Early poisoning attacks targeted support vector machines and simple neural networks (Biggio et al., 2012; Koh & Liang, 2017). As poisoning gained popularity, various strategies for triggerless attacks on deep architectures emerged (Muñoz-González et al., 2017; Shafahi et al., 2018; Zhu et al., 2019; Huang et al., 2020; Aghakhani et al., 2020b; Geiping et al., 2020). The early backdoor attacks contained triggers in the poisoned data and in some cases changed the label, thus were not clean-label (Chen et al., 2017; Gu et al., 2017; Liu et al., 2017; Salem et al., 2020b; Yao et al., 2019). However, methods that produce poison examples which do not visibly contain a trigger also show positive

results (Chen et al., 2017; Turner et al., 2018; Saha et al., 2019; Salem et al., 2020a). Poisoning attacks have also precipitated several defense strategies, but sanitization-based defenses may be overwhelmed by some attacks (Koh et al., 2018; Liu et al., 2018; Chacon et al., 2019; Peri et al., 2020).

We focus on attacks that achieve targeted misclassification. That is, under both the triggerless and backdoor threat models, the end goal of an attacker is to cause a target sample to be misclassified as another specified class. Other objectives, such as decreasing overall test accuracy, have been studied, but less work exists on this topic with respect to neural networks (Xiao et al., 2015; Liu et al., 2020). In both triggerless and backdoor data poisoning, the clean images, called *base images*, that are modified by an attacker come from a single class, the *base class*. This class is often chosen to be precisely the same class into which the attacker wants the target image or class to be misclassified.

There are two major differences between triggerless and backdoor threat models in the literature. First and foremost, backdoor attacks alter their targets during inference by adding a trigger. In the works we consider, these triggers take the form of small patches added to an image (Turner et al., 2018; Saha et al., 2019). Second, these works on backdoor attacks cause a victim to misclassify any image containing the trigger rather than a particular sample. Triggerless attacks instead cause the victim to misclassify an individual image called the *target image* (Shafahi et al., 2018; Zhu et al., 2019; Aghakhani et al., 2020b; Geiping et al., 2020). This second distinction between the two threat models is not essential; for example, triggerless attacks could be designed to cause the victim to misclassify a collection of images rather than a single target. To be consistent with the literature at large, we focus on triggerless attacks that target individual samples and backdoor attacks that target whole classes of images.

We focus on the *clean-label backdoor attack* and the *hidden trigger backdoor attack*, where poisons are crafted with optimization procedures and do not contain noticeable patches (Saha et al., 2019; Turner et al., 2018). For triggerless attacks, we focus on the *feature collision* and *convex polytope* methods, the most highly cited attacks of the last two years that have appeared at prominent ML conferences (Shafahi et al., 2018; Zhu et al., 2019). We include the recent triggerless methods *Bullseye Polytope* (BP) and *Witches’ Brew* (WiB) in the section where we present metrics on our benchmark problems (Aghakhani et al., 2020b; Geiping et al., 2020). The following section details the attacks that serve as the subjects of our experiments.

**Technical details:** Before formally describing various poisoning methods, we begin with notation. Let  $X_c$  be the set of all clean training data, and let  $X_p = \{x_p^{(j)}\}_{j=1}^J$

denote the set of  $J$  poison examples with corresponding clean base images  $\{x_b^{(j)}\}_{j=1}^J$ . Let  $x_t$  be the target image. Labels are denoted by  $y$  and  $Y$  for a single image and a set of images, respectively, and are indexed to match the data. We use  $f$  to denote a feature extractor network.

**Feature Collision (FC)** Poisons in this attack are crafted by adding small perturbations to base images so that their feature representations lie extremely close to that of the target (Shafahi et al., 2018). Formally, each poison is the solution to the following optimization problem.

$$x_p^{(j)} = \operatorname{argmin}_x \|f(x) - f(x_t)\|_2^2 + \beta \|x - x_b^{(j)}\|_2^2. \quad (1)$$

When we enforce  $\ell_\infty$ -norm constraints, we drop the last term in Equation (1) and instead enforce  $\|x_p^{(j)} - x_b^{(j)}\|_\infty \leq \varepsilon$ ,  $\forall j$  by projecting onto the  $\ell_\infty$  ball after each iteration of the optimization procedure.

**Convex Polytope (CP)** This attack crafts poisons such that the target’s feature representation is a convex combination of the poisons’ feature representations by solving the following optimization problem (Zhu et al., 2019).

$$\begin{aligned} X_p = \operatorname{argmin}_{\{c_j\}, \{x^{(j)}\}} & \quad \frac{1}{2} \frac{\|f(x_t) - \sum_{j=1}^J c_j f(x^{(j)})\|_2^2}{\|f(x_t)\|_2^2} \\ \text{subject to} & \quad \sum_{j=1}^J c_j = 1 \\ & \quad \text{and } c_j \geq 0 \forall j, \\ & \quad \text{and } \|x^{(j)} - x_b^{(j)}\|_\infty \leq \varepsilon \forall j. \end{aligned} \quad (2)$$

**Clean Label Backdoor (CLBD)** This backdoor attack begins by computing an adversarial perturbation to each base image (Turner et al., 2018). Formally,

$$\hat{x}_p^{(j)} = x_b^{(j)} + \operatorname{argmax}_{\|\delta\|_\infty \leq \varepsilon} \mathcal{L}(x_b^{(j)} + \delta, y^{(j)}; \theta), \quad (3)$$

where  $\mathcal{L}$  denotes cross-entropy loss. Then, a patch is added to each image in  $\{\hat{x}_p^{(j)}\}$  to generate the final poisons  $\{x_p^{(j)}\}$ . The patched image is subject to an  $\ell_\infty$ -norm constraint.

**Hidden Trigger Backdoor (HTBD)** A backdoor analogue of the FC attack, where poisons are crafted to remain close to the base images but collide in feature space with a patched image from the target class (Saha et al., 2019). Let  $\tilde{x}_t^{(j)}$  denote a patched training image from the target class (this image is not clean), then we solve the following optimization problem to find poison images;

$$\begin{aligned} x_p^{(j)} = \operatorname{argmin}_x & \quad \|f(x) - f(\tilde{x}_t^{(j)})\|_2^2, \\ \text{subject to} & \quad \|x - x_b^{(j)}\|_\infty \leq \varepsilon. \end{aligned} \quad (4)$$

### 3. Why Do We Need Benchmarks?

Backdoor and triggerless attacks have been tested in a wide range of disparate settings. From model architecture to target/base class pairs, the literature is inconsistent. Experiments are also lacking in the breadth of trials performed, sometimes using only one model initialization for all experiments, or testing against one single target image. We find that inconsistencies in experimental settings have a large impact on performance evaluations and have resulted in comparisons that are difficult to interpret. For example, the authors of CP compare their  $\ell_\infty$ -constrained attack to FC, which is crafted with an  $\ell_2$  penalty. In other words, these methods have never been compared on a level playing field.

To study these attacks thoroughly and rigorously, we employ sampling techniques that allow us to draw conclusions about the attacks taking into account variance across model initializations and class choice. For a single trial, we sample one of ten checkpoints of a given architecture, then randomly select the target image, base class, and base images. In Section 4, all figures are averages from 100 trials with our sampling techniques.

**Disparate evaluation settings in the literature.** To understand how differences in evaluation settings impact results, we re-create the various original performance tests for each of the methods described above within our common evaluation framework. We try to be as faithful as possible to the original works, however we employ our own sampling techniques described above to increase statistical significance. Then, we tweak these experiments one component at a time revealing the fragility of each method to changes in evaluation setup. While proof-of-concept papers that propose novel methods have great value in furthering the community’s understanding of the threat posed by large unchecked datasets, comparing strategies on the same task and comparing their sensitivity to experimental design changes are vital too. The variations in experimental design for the most part do not correspond to differences in threat models or in adversarial goals, and where they do, like transfer learning versus training from scratch, performance across the board may be hard to predict and thus requires careful examination.

**Establishing baselines.** For the FC setting, following one of the main setups in the original paper, we craft 50 poisons on an AlexNet variant (for details on the specific architecture, see (Krizhevsky et al., 2012; Shafahi et al., 2018)) pre-trained on CIFAR-10 (Krizhevsky et al., 2009), and we use the  $\ell_2$ -norm penalty version of the attack. We then evaluate poisons on the same AlexNet, using the same CIFAR-10 data to train for 20 more epochs to “fine tune” the model end to end. Note that this is not really transfer learning in the usual sense, as the fine tuning utilizes the same dataset

Table 1. Various experimental designs used in data poisoning research.

Attack	Data		Opt.	Transfer Learning			Threat Model			Ensembles	$\epsilon$
	Norm.	Aug.	SGD	FFE	E2E	FST	WB	GB	BB		
FC	×	×	×	✓	✓	×	✓	×	×	×	-
CP	✓	×	×	✓	✓	×	×	✓	✓	✓	25.5
CLBD	×	✓	✓	×	×	✓	×	×	✓	×	8
HTBD	✓	×	✓	✓	×	×	✓	×	×	×	16

as pre-training, except with poisons inserted (Shafahi et al., 2018).

The CP setting involves crafting five poisons using a ResNet-18 model (He et al., 2016) pre-trained on CIFAR-10, and then fine tuning the linear layer of the same ResNet-18 model with a subset of the CIFAR-10 training comprising 50 images per class (including the poisons). This setup is also not representative of typical transfer learning, as the fine-tuning data is sub-sampled from the pre-training dataset. In this baseline we set  $\epsilon = 25.5/255$  matching the original work (Zhu et al., 2019).

One of the original evaluation settings for CLBD uses 500 poisons. We craft these on an adversarially trained ResNet-18 and modify them with a  $3 \times 3$  patch in the lower right-hand corner. The perturbations are bounded with  $\epsilon = 16/255$ . We then train a narrow ResNet model from scratch with the CIFAR-10 training set (including the poisons) (Turner et al., 2018).

For the HTBD setting, we generate 800 poisons with another modified AlexNet (for architectural details, see Appendix A.13) which is pre-trained on CIFAR-10 dataset. Then, an  $8 \times 8$  trigger patch is added to the lower right corner of the target image, and the perturbations are bounded with  $\epsilon = 16/255$ . We use the entire CIFAR-10 dataset (including the poisons) to fine tune the last fully connected layer of the same model used for crafting. Once again, the fine-tuning data in this setup is not disjoint from the pre-training data (Saha et al., 2019). See Table 2 and the left-most bars of Figure 3 for all baseline results.

**Inconsistencies in previous work.** The baselines defined above do not serve as a fair comparison across methods, since the original works to which we try and stay faithful are inconsistent. Table 1 summarizes experimental settings in the original works. If a particular component (column header) was considered anywhere in the original paper’s experiments, we mark a (✓), leaving exes (×) when something was not present in any experiments. Table 1 shows the presence of data normalization and augmentation as well as optimizers (SGD or ADAM). It also shows which learning setup the original works considered: frozen feature extractor (FFE), end-to-end fine tuning (E2E), or from-scratch

training (FST), as well as which threat levels were tested, white, grey or black box (WB, GB, BB). We also consider whether or not an ensembled attack was used. The  $\epsilon$  values reported are out of 255 and represent the smallest bound considered for CIFAR-10 poisons in the papers; note FC uses an  $\ell_2$  penalty so no bound is enforced despite the attack being called “clean-label” in the original work. We conclude from Table 1 that experimental design varies greatly from paper to paper, making it extremely difficult to make any comparisons between methods.

#### 4. Just How Toxic Are Poisoning Methods Really?

In this section, we look at weaknesses and inconsistencies in existing experimental setups, and how these lead to potentially misleading comparisons between methods. We use our testing framework to put triggerless and backdoor attacks to the test under a variety of circumstances, and get a tighter grip on the reliability of existing poisoning methods.

Table 2. Baseline performance.

Attack	Success Rate (%)
FC	92.00 ± 2.71
CP	88.00 ± 3.25
CLBD	86.00 ± 3.47
HTBD	69.00 ± 4.62

**Training with SGD and data augmentation.** In their corresponding original works, both FC and CP attacks have only been tested on victim models pre-trained with the ADAM optimizer. However, SGD with momentum has become the dominant optimizer for training CNNs (Wilson et al., 2017). Interestingly, we find that models trained with SGD are significantly harder to poison, rendering these attacks less effective in practical settings. Moreover, none of the baselines include simple data augmentation such as horizontal flips and random crops. We find that data augmentation, standard in the deep learning literature, also greatly reduces the effectiveness of all of the attacks. For example, FC and CP success rates plummet in this setting to 51.00% and 19.09%, respectively. Complete results including hy-

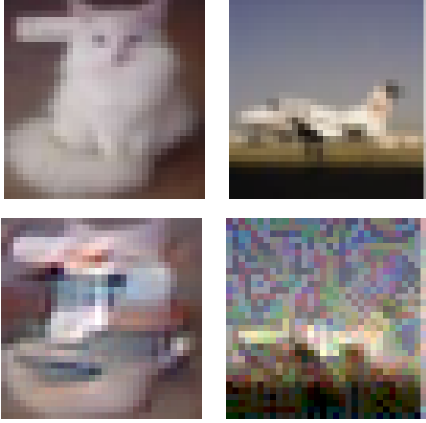


Figure 1. Bases (top) and poisons (bottom).

perparameters, success rates, and confidence intervals are reported in Appendix A.3.

**Victim architecture matters.** Two attacks, FC and HTBD, are originally tested on AlexNet variants, and CLBD is tested with a narrow ResNet variant. These models are not widely used, and they are unlikely to be employed by a realistic victim. We observe that many attacks are significantly less effective against ResNet-18 victims. See Figure 3, where for example, the success rate of HTBD on these victims is as low as 18%. See Appendix A.4 for a table of numerical results. These ablation studies are conducted in the baseline settings but with a ResNet-18 victim architecture. These ResNet experiments serve as an example of how performance can be highly dependent on the selection of architecture.

**“Clean” attacks are sometimes dirty.** Each of the original works we consider purports to produce “clean-label” poison examples that look like natural images. However these methods often produce easily visible image artifacts and distortions due to the large values of  $\epsilon$  used. See Figure 1 for examples generated by two of the methods, where FC perturbs a clean “cat” into an unrecognizable poison (left), and CP generates an extremely noisy poison from a base in the “airplane” class (right). These images are not surprising since the FC method is tested with an  $\ell_2$  penalty in the original work, and CP is  $\ell_\infty$  constrained with a large radius of  $25.5/255$ .

In many contexts, avoiding detection by automated systems may be more important than maintaining perceptual similarity. In our work, we focus on perceptual similarity as defined by the  $\ell_\infty$  constraint as this reflects the explicit goal of most of the attacks we examine, and it is, in general, a much more common area of study. Adaptive attacks that avoid defense or detection is relatively unexplored and an

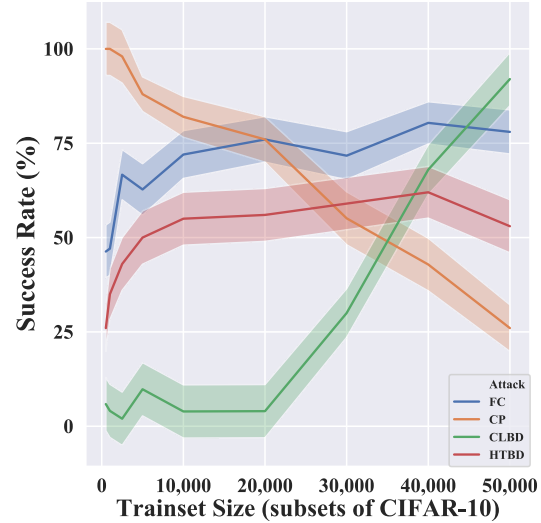


Figure 2. Scaling the dataset size while fixing the poison budget.

interesting area for future research (Koh et al., 2018).

Borrowing from common practice in the evasion attack and defense literature, we test each method with an  $\ell_\infty$  constraint of radius  $8/255$  and find that the effectiveness of every attack is diminished (Madry et al., 2017; Dong et al., 2020). The sensitivity to perturbation size suggests that a standardized constraint on poison examples is necessary for fair comparison of attacks. See Figure 3, and see Appendix A.5 for a table of numerical results.

**Proper transfer learning may be less vulnerable.** Of the attacks we study here, FC, CP, and HTBD were originally proposed in settings referred to as “transfer learning.” Each particular setup varies, but none are true transfer learning since the pre-training and fine-tuning datasets overlap. For example, FC uses the entire CIFAR-10 training dataset for both pre-training and fine tuning. Thus, their threat model entails allowing an adversary to modify the training dataset but only for the last few epochs. Furthermore, these attacks use inconsistently sized fine-tuning datasets.

To simulate transfer learning, we test each attack with ResNet-18 feature extractors pre-trained on CIFAR-100, which are then fine tuned on CIFAR-10 data. In Figure 3, every attack aside from CP shows worse performance when transfer learning is done on data that is disjoint from the pre-training dataset. The attacks designed for transfer learning may not work as advertised in more realistic transfer learning settings. See Appendix A.6.

**Performance is not invariant to dataset size.** Existing work on data poisoning measures an attacker’s budget in terms of what percentage of the training data they may

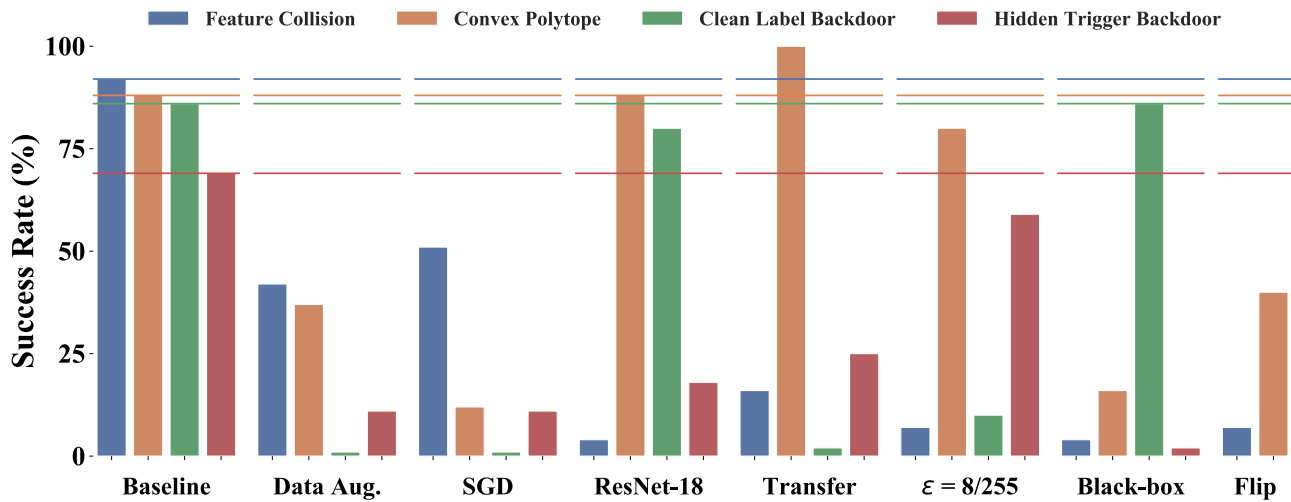


Figure 3. We show the fragility of poisoning methods to experimental design. This figure depicts baselines along with the results of ablation studies. Different methods respond differently to these testing scenarios, supporting the need for consistent and thorough testing. Horizontal lines denote performance on baselines described in Section 3, and bars represent the results of changing a specific feature in an individual method’s baseline. Tables of these results with confidence intervals can be found in the appendices.

modify. This begs the question whether percentage alone is enough to characterize the budget. Does the actual size of the training set matter? We find the number of images in the training set has a large impact on attack performance, and that performance curves for FC and CP intersect. When we hold the percentage poisoned constant at 1%, but we change the number of poisons, and the size of the training set accordingly, we see no consistent trends in how the attacks are affected. Figure 2 shows the success of each attack as a function of dataset size (shaded region is one standard error). This observation suggests that one cannot compare attacks tested on different sized datasets by only fixing the percent of the dataset poisoned. See Appendix A.7.

**Black-box performance is low.** Whether considering transfer learning or training from scratch, testing these methods against a black-box victim is surely one of the most realistic tests of the threat they pose. Since, FC, CP and HTBD do not consider the black-box scenario in the original works, we take the poisons crafted using baseline methods and evaluate them on models of different architectures than those used for crafting. The attacks show much lower performance in the black-box settings than in the baselines, in particular FC, CP, and HTBD all have success rates lower than 20%. See Figure 3, and see Appendix A.8 for more details.

**Small sample sizes and non-random targets.** On top of inconsistencies in experimental setups, existing work on data poisoning often test only on specific target/base class pairs. For example, FC largely uses “frog” as the base class and “airplane” as the target class. CP, on the other hand, only uses “ship” and “frog” as the base and target classes, respectively. Neither work contains experiments where each trial consists of a randomly selected target/base class pair. We find that the success rates are highly class pair dependent and change dramatically under random class pair sampling. For this reason, random sampling of image pairs is a good step towards achieving consistent and reproducible results. See Appendix A.9 for a comparison of the specific class pairs from these original works with randomly sampled class pairs.

In addition to inconsistent class pairs, data poisoning papers often evaluate performance with very few trials since the methods are computationally expensive. In their original works, FC and CP use 30 and 50 trials, respectively, for each experiment, and these experiments are performed on the same exact pre-trained models each time. And while HTBD does test randomized pairs, they only show results for ten trials on CIFAR-10. These small sample sizes yield wide error bars in performance evaluation. We choose to run 100 trials per experiment in our own work. While we acknowledge that a larger number would be even more compelling, 100 is a compromise between thorough experimentation and practicality since each trial requires re-training a classifier.

**Attacks are highly specific to the target image.** Triggerless attacks have been proposed as a threat against systems deployed in the physical world. For example, blue Toyota sedans may go undetected by a poisoned system so that an attacker may fly under the radar. However, triggerless attacks are generally crafted against a specific target image, while a physical object may appear differently under different real-world circumstances. We upper-bound the robustness of poison attacks by applying simple horizontal flips to the target images, and we find that these poisoning methods are weak when the exact target image is unknown. For example, FC is only successful 7% of the time when simply flipping the target image. See Figure 3 and Appendix A.10.

**Backdoor success depends on patch size.** Backdoor attacks add a patch to target images to trigger misclassification. In real-world scenarios, a small patch may be critical to avoid being caught. The original HTBD attack uses an  $8 \times 8$  patch, while the CLBD attack originally uses a  $3 \times 3$  patch (Saha et al., 2019; Turner et al., 2018). In order to understand the impact on attack performance, we test different patch sizes. We find a strong correlation between the patch size and attack performance, see Appendix A.12. We conclude that backdoor attacks must be compared using identical patch sizes.

## 5. Evaluation Metrics for Dataset Manipulation

**Our Benchmark:** We propose new benchmarks for measuring the efficacy of *both* backdoor and triggerless data poisoning attacks. The deviations from the original settings in which methods were proposed are carefully chosen to keep these benchmark tasks in line with the original threats while leveling the playing field for fair comparison. We standardize the datasets and problem settings for our benchmarks as described below.<sup>1</sup>

Target and base images are chosen from the testing and training sets, respectively, according to a seeded/reproducible random assignment. Poison examples crafted from the bases must remain within the  $\ell_\infty$ -ball of radius  $8/255$  centered at the corresponding base images. Seeding the random assignment allows us to test against a significant number of different random choices of base/target, while always using the same choices for each method, thus removing a source of variation from the results. We consider two different training modes:

**I Transfer Learning:** A feature extractor pre-trained on clean data is frozen and used while training a linear

classification head on a disjoint set of training data that contains poisons.

**II Training From Scratch:** A network is trained from random initialization on data containing poison examples in the training set.

To further standardize these tests, we provide pre-trained models to test against. The parameters of one model are given to the attacker. We then evaluate the strength of the attacks in white-box and black-box scenarios. For white-box tests in the transfer learning benchmarks, we use the same frozen feature extractor that is given to the attacker for evaluation. While in the black-box setting, we craft poisons using the known model but we test on the two models the attacker has not seen, averaging the results. When training from scratch, models are trained from a random initialization on the poisoned dataset. We report averages from 100 independent trials for each test. Backdoor attacks can use any  $5 \times 5$  patch. Note that the number of attacker-victim network pairs is kept small in our benchmark because each of the 100 trials requires re-training (in some cases from scratch), and we want to keep the benchmark within reach for researchers with modest computing resources.

**CIFAR-10 benchmarks.** Models are pretrained on CIFAR-100, and the fine-tuning data is a subset of CIFAR-10. We choose this subset to be the first 250 images per class (2,500 images), this includes 25 poison examples in total (2,475 unperturbed images). This amount of data motivates the use of transfer learning, since training from scratch on only 2,500 images yields poor generalization. See Appendix A.13 for examples. We allow 500 poisons when training from scratch, see Appendix A.15 for a case-study in which we investigate how many poisons an attacker may be able to place in a dataset compiled by querying the internet for images. We allow the attacker access to a ResNet-18, and we do black-box tests on a VGG11 (Simonyan & Zisserman, 2014), and a MobileNetV2 (Sandler et al., 2018), and we use one of each model when training from scratch and report the average.

**TinyImageNet benchmarks.** Additionally, we pre-train VGG16, ResNet-34, and MobileNetV2 models on the first 100 classes of the TinyImageNet dataset (Le & Yang, 2015). We fine tune these models on the second half of the dataset, allowing for 250 poison images. As above, the attacker has access to a particular VGG16 model, and black-box tests are done on the other two models. In the from-scratch setting, we train a VGG16 model on the entire TinyImageNet dataset with 250 images poisoned.<sup>2</sup>

<sup>1</sup>Code is available at <https://github.com/aks2203/poisoning-benchmark>.

<sup>2</sup>The TinyImageNet from-scratch benchmark is done with 25 independent trials to keep the computational demands this problem within reach for researchers with modest resources.

Table 3. Benchmark success rates (reported as percentages). The best performance in each column is in bold.

Attack	CIFAR-10			TinyImageNet		
	Transfer		From Scratch	Transfer		From Scratch
	WB	BB		WB	BB	
FC	22.0	7.0	1.33	49.0	2.0	4.0
CP	33.0	7.0	0.67	14.0	1.0	0.0
BP	<b>85.0</b>	8.5	2.33	<b>100.0</b>	<b>10.5</b>	<b>44.0</b>
WiB	-	-	<b>26.0</b>	-	-	32.0
CLBD	5.0	6.5	1.00	3.0	1.0	0.0
HTBD	10.0	<b>9.5</b>	2.67	3.0	0.5	0.0

**Benchmark hyperparameters** We pre-train models on CIFAR-100 with SGD for 400 epochs starting with a learning rate of 0.1, which decays by a factor of 10 after epochs 200, 300, and 350. Models pre-trained on the first half of TinyImageNet are trained with SGD for 200 epochs starting with a learning rate of 0.1, which decays by a factor of 10 after epochs 100 and 150. In both cases, we apply per-channel data normalization, random crops, and horizontal flips, and we use batches of 128 images (augmentation is also applied to the poisoned images). We then fine tune with poisoned data for 40 epochs with a learning rate that starts at 0.01 and drops to 0.001 after the 30<sup>th</sup> epoch (this applies to the transfer learning settings).

When training from scratch on CIFAR-10, we include the 500 perturbed poisons in the standard training set. We use SGD and train for 200 epochs with batches of 128 images and an initial learning rate of 0.1 that decays by a factor of 10 after epochs 100 and 150. Here too, we use data normalization and augmentation as described above. When training from scratch on TinyImageNet, we allow for 250 poisoned images. All other hyperparameters are identical.

Our evaluations of six different attacks are shown in Table 3. These attacks are not easily ranked, as the strongest attacks in some settings are not the strongest in others. Witches’ Brew (WiB) is not evaluated in the transfer learning settings, since it is not considered in the original work (Geiping et al., 2020). See Appendix A.16 for tables with confidence intervals. We find that by using disjoint and standardized datasets for transfer learning, and common training practices like data normalization and scheduled learning rate decay, we overcome the deficits in previous work. Our benchmarks can provide useful evaluations of data poisoning methods and meaningful comparisons between them.

## 6. Conclusion

The threat of data poisoning is at the forefront of fears around emerging ML systems (Kumar et al., 2020a). While

many of the methods claiming to do so do not pose a practical threat, some of the recent methods are cause for practitioner concern. With real threats arising, there is a need for fair comparison. The diversity of attacks, and in particular the difficulty in ordering them by efficacy, calls for a diverse set of benchmarks. With those we present here, practitioners and researchers can compare attacks on a level playing field and gain an understanding of how existing methods match up with one another and where they might fail.

Since the future advancement of these methods is inevitable, our benchmarks will also serve the data poisoning community as a standardized test problem on which to evaluate and future attack methodologies. As even stronger attacks emerge, trepidation on the part of practitioners will be matched by the potential harm of poisoning attacks. We are arming the community with the high quality metrics this evolving situation calls for.

## Acknowledgements

This work was supported by DARPA GARD, the AFOSR MURI program, the Office of Naval Research, and the DARPA YFA program.

## References

Aghakhani, H., Eisenhofer, T., Schönherr, L., Kolossa, D., Holz, T., Kruegel, C., and Vigna, G. Venomave: Clean-label poisoning against speech recognition. *arXiv preprint arXiv:2010.10682*, 2020a.

Aghakhani, H., Meng, D., Wang, Y.-X., Kruegel, C., and Vigna, G. Bullseye polytope: A scalable clean-label poisoning attack with improved transferability. *arXiv preprint arXiv:2005.00191*, 2020b.

Biggio, B., Nelson, B., and Laskov, P. Poisoning attacks against support vector machines. In *Proceedings of the*



- 29th International Conference on International Conference on Machine Learning, ICML'12, pp. 1467–1474, 2012.
- Chacon, H., Silva, S., and Rad, P. Deep learning poison data attack detection. In *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 971–978, 2019.
- Chen, X., Liu, C., Li, B., Lu, K., and Song, D. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.
- Dai, J., Chen, C., and Li, Y. A backdoor attack against lstm-based text classification systems. *IEEE Access*, 7: 138872–138878, 2019.
- Dong, Y., Fu, Q.-A., Yang, X., Pang, T., Su, H., Xiao, Z., and Zhu, J. Benchmarking adversarial robustness on image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 321–331, 2020.
- Fang, M., Yang, G., Gong, N. Z., and Liu, J. Poisoning attacks to graph-based recommender systems. In *Proceedings of the 34th Annual Computer Security Applications Conference*, pp. 381–392, 2018.
- Fang, M., Gong, N. Z., and Liu, J. Influence function based data poisoning attacks to top-n recommender systems. In *Proceedings of The Web Conference 2020*, pp. 3019–3025, 2020.
- Geiping, J., Fowl, L., Huang, W. R., Czaja, W., Taylor, G., Moeller, M., and Goldstein, T. Witches’ brew: Industrial scale data poisoning via gradient matching. *arXiv preprint arXiv:2009.02276*, 2020.
- Gu, T., Dolan-Gavitt, B., and Garg, S. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017.
- 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.
- Hu, R., Guo, Y., Pan, M., and Gong, Y. Targeted poisoning attacks on social recommender systems. In *2019 IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6. IEEE, 2019.
- Huang, W. R., Geiping, J., Fowl, L., Taylor, G., and Goldstein, T. Metapoinson: Practical general-purpose clean-label data poisoning. *Advances in Neural Information Processing Systems*, 33, 2020.
- Jiang, L., Zhou, Z., Leung, T., Li, L.-J., and Fei-Fei, L. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. *arXiv preprint arXiv:1712.05055*, 2017.
- Koh, P. W. and Liang, P. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 1885–1894. JMLR. org, 2017.
- Koh, P. W., Steinhardt, J., and Liang, P. Stronger data poisoning attacks break data sanitization defenses. *arXiv preprint arXiv:1811.00741*, 2018.
- Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.
- Krizhevsky, A., Sutskever, I., and Hinton, G. E. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.
- Kumar, R. S. S., Nyström, M., Lambert, J., Marshall, A., Goertzel, M., Comissoneru, A., Swann, M., and Xia, S. Adversarial machine learning-industry perspectives. In *2020 IEEE Security and Privacy Workshops (SPW)*, pp. 69–75. IEEE, 2020a.
- Kumar, R. S. S., Nyström, M., Lambert, J., Marshall, A., Goertzel, M., Comissoneru, A., Swann, M., and Xia, S. Adversarial machine learning-industry perspectives. *arXiv preprint arXiv:2002.05646*, 2020b.
- Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Mallocci, M., Kolesnikov, A., et al. The open images dataset v4. *International Journal of Computer Vision*, pp. 1–26, 2020.
- Le, Y. and Yang, X. Tiny imagenet visual recognition challenge. *CS 231N*, 7, 2015.
- Li, B., Wang, Y., Singh, A., and Vorobeychik, Y. Data poisoning attacks on factorization-based collaborative filtering. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 1893–1901, 2016.
- Liu, K., Dolan-Gavitt, B., and Garg, S. Fine-pruning: Defending against backdooring attacks on deep neural networks. In *International Symposium on Research in Attacks, Intrusions, and Defenses*, pp. 273–294. Springer, 2018.
- Liu, S., Lu, S., Chen, X., Feng, Y., Xu, K., Al-Dujaili, A., Hong, M., and O’Reilly, U.-M. Min-max optimization without gradients: Convergence and applications to black-box evasion and poisoning attacks. In *International Conference on Machine Learning*, pp. 6282–6293. PMLR, 2020.

- Liu, Y., Ma, S., Aafer, Y., Lee, W.-C., Zhai, J., Wang, W., and Zhang, X. Trojaning attack on neural networks. 2017.
- Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.
- Muñoz-González, L., Biggio, B., Demontis, A., Paudice, A., Wongrassamee, V., Lupu, E. C., and Roli, F. Towards poisoning of deep learning algorithms with back-gradient optimization. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pp. 27–38. ACM, 2017.
- Ni, J., Li, J., and McAuley, J. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 188–197, 2019.
- Peri, N., Gupta, N., Huang, W. R., Fowl, L., Zhu, C., Feizi, S., Goldstein, T., and Dickerson, J. P. Deep k-nn defense against clean-label data poisoning attacks. In *European Conference on Computer Vision*, pp. 55–70. Springer, 2020.
- Saha, A., Subramanya, A., and Pirsiavash, H. Hidden trigger backdoor attacks. *arXiv preprint arXiv:1910.00033*, 2019.
- Salem, A., Backes, M., and Zhang, Y. Don’t trigger me! a triggerless backdoor attack against deep neural networks. *arXiv preprint arXiv:2010.03282*, 2020a.
- Salem, A., Wen, R., Backes, M., Ma, S., and Zhang, Y. Dynamic backdoor attacks against machine learning models. *arXiv preprint arXiv:2003.03675*, 2020b.
- Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.
- Shafahi, A., Huang, W. R., Najibi, M., Suci, O., Studer, C., Dumitras, T., and Goldstein, T. Poison frogs! targeted clean-label poisoning attacks on neural networks. In *Advances in Neural Information Processing Systems*, pp. 6103–6113, 2018.
- Simonyan, K. and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
- Solans, D., Biggio, B., and Castillo, C. Poisoning attacks on algorithmic fairness. *arXiv preprint arXiv:2004.07401*, 2020.
- Turner, A., Tsipras, D., and Madry, A. Clean-label backdoor attacks. 2018.
- Wilson, A. C., Roelofs, R., Stern, M., Srebro, N., and Recht, B. The marginal value of adaptive gradient methods in machine learning. In *Advances in neural information processing systems*, pp. 4148–4158, 2017.
- Xiao, H., Biggio, B., Nelson, B., Xiao, H., Eckert, C., and Roli, F. Support vector machines under adversarial label contamination. *Neurocomputing*, 160:53–62, 2015.
- Yao, Y., Li, H., Zheng, H., and Zhao, B. Y. Latent backdoor attacks on deep neural networks. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pp. 2041–2055, 2019.
- Zhu, C., Huang, W. R., Li, H., Taylor, G., Studer, C., and Goldstein, T. Transferable clean-label poisoning attacks on deep neural nets. In *International Conference on Machine Learning*, pp. 7614–7623, 2019.