DP-InstaHide: Provably Defusing Poisoning and Backdoor Attacks with Differentially Private Data Augmentations

Eitan Borgnia
Department of Computer Science
University of Maryland
eborgnia2@gmail.com

Jonas Geiping
Department of EE & Computer Science
University of Siegen
jonas.geiping@uni-siegen.de

Valeriia Cherepanova
Department of Computer Science
University of Maryland
vcherepa@umd.edu

Liam Fowl
Department of Mathematics
University of Maryland
lfowl@math.umd.edu

Arjun Gupta
Department of Computer Science
University of Maryland
arjung15@umd.edu

Amin Ghiasi
Department of Computer Science
University of Maryland
amin@cs.umd.edu

Furong Huang
Department of Computer Science
University of Maryland
furongh@umd.edu

Micah Goldblum
Department of Computer Science
University of Maryland
goldblumcello@gmail.com

Tom Goldstein
Department of Computer Science
University of Maryland
tomg@cs.umd.edu

Abstract

Data poisoning and backdoor attacks manipulate training data to induce security breaches in a victim model. Such attacks can be provably deflected using differentially private (DP) training methods, though these come with a sharp decrease in model performance. InstaHide has recently been proposed as an alternative to DP training, which leverages supposed privacy properties of mixup augmentation but lacks rigorous guarantees. We offer a modification of this method, DP-InstaHide, which works by combining the mixup regularizer with additive noise. A rigorous analysis of DP-InstaHide shows that mixup does indeed have privacy advantages and training with k-way mixup provably yields at least k times stronger DP guarantees than a naive DP mechanism. Because mixup (as opposed to noise) is beneficial to model performance, DP-InstaHide provides a mechanism for achieving stronger empirical performance against poisoning attacks than other known DP methods. Moreover, we empirically verify that strong data augmentations, such as mixup and random additive noise, indeed nullify poison attacks while enduring only a small accuracy trade-off.

1 Introduction

As the capabilities of machine learning systems expand, so do their training data demands. To satisfy this massive data requirement, developers create automated web scrapers that download data without human supervision. The lack of human control over the machine learning pipeline may expose systems to *poisoned* training data that induces pathologies in models trained on it. Data poisoning and backdoor attacks may degrade accuracy or elicit incorrect predictions in the presence of a triggering visual feature (Shafahi et al., 2018; Chen et al., 2017).

To combat this threat model, a number of defenses against data poisoning have emerged. Certified defenses based on *differential privacy* (DP) provably desensitize models to small changes in their training data by adding noise to either the data or the gradients used by their optimizer (Ma et al., 2019). When a model is trained using sufficiently strong DP, it is not possible to infer whether a small collection of data points were present in the training set by observing model behaviors, and it is therefore not possible to significantly alter model behaviors by introducing a small number of poisoned samples. In this work, we show that strong data augmentations, specifically mixup Zhang et al. (2017) and its variants, provide state-of-the-art empirical defense against data poisoning, backdoor attacks, and even adaptive attacks. This good performance can be explained by the differential privacy benefits of mixup.

We present a variant of InstaHide (Huang et al., 2020b) with rigorous privacy guarantees and study its use to rebuff poisoning attacks. Like the original InstaHide, our approach begins by applying mixup augmention to a dataset. However, we do not use the random multiplicative mask and instead introduce randomness via added Laplacian noise. Our approach exploits the fact that mixup augmentation concentrates training data near the center of the ambient unit hypercube and saturates this region of space more densely than the original dataset. Hence, less noise is required to render the data private than if noise were added to the original data. In fact, we show that adding noise on top of k-way mixup creates a differential privacy guarantee that is k times stronger (i.e., ϵ is k time smaller) than adding noise alone.

In addition to mixup, we also perform experiments with the related CutMix and MaxUp augmentations. Because these augmentations are designed for improving generalization in image classifiers, we find that they yield a favorable robustness accuracy trade-off compared to other strong defenses (Yun et al., 2019; Zhang et al., 2017; Gong et al., 2020).

1.1 Related Work

Broadly speaking, data poisoning attacks aim to compromise the performance of a network by maliciously modifying the data on which the network is trained. In this paper, we examine three classes of such attacks:

Backdoor attacks involve inserting a "trigger," often a fixed patch, into training data. Attackers can then add the same patch to data at test time to fool the network into misclassifying modified images as the target class (Gu et al., 2017; Tran et al., 2018b; Saha et al., 2020). *Feature collision* attacks occur when the attacker modifies training samples so they collide with, or surround, a target test-time image in feature space (Shafahi et al., 2018; Zhu et al., 2019; Aghakhani et al., 2020). These methods work primarily in the transfer learning setting, where a known feature extractor is fixed and a classification layer is fine-tuned on the perturbed data. *From-scratch* attacks modify training data to cause targeted misclassification of pre-selected test time images. Crucially, these attacks work in situations where a deep network is *a priori* trained on modified data, rather than being pre-trained and subsequently fine-tuned on poisoned data (Huang et al., 2020a; Geiping et al., 2020).

A variety of defenses against poisoning attacks have also been proposed:

Filtering defenses, which either remove or relabel poisoned data, are the most common type of defense for targeted attacks. These methods rely on the tendency of poisoned data to differ sufficiently from clean data in feature space. *Differentially private SGD* is a principled defense, where training gradients are clipped and noised, thus diminishing the effects of poisoned gradient updates. However, these defenses have been shown to fail against advanced attacks, as they often lead to significant drops in clean validation accuracy Geiping et al. (2020).

Outside of data poisoning, Lee et al. (2019) introduce *DP-Mix*, which connects data augmentation and privacy by using tools for Rényi differential privacy for subsampling (Wang et al., 2019) to analyze Rényi bounds for image mixtures with Gaussian noise. While these bounds can readily be converted into differential privacy guarantees, they suffer from numeric instability and tend to be loose in the low privacy regime, where validation accuracy is maintained. *InstaHide*, proposed by

Huang et al. (2020b), uses mixup combined with a random mask to achieve dataset privacy, but was found to fail by Carlini et al. (2020).

2 DP-InstaHide: A Mixup Defense with Provable Differential Privacy Advantages

The original InstaHide method attempted to privatize data by first applying mixup, and then multiplying the results by random binary masks. While the idea that mixup enhances the privacy of a dataset is well founded, the original InstaHide scheme lies outside of the classical differential privacy framework and is now known to be insecure. We propose a variant of the method, DP-InstaHide, which replaces the multiplicative random mask with additive random noise (see Figure A.3). The resulting method comes with a differential privacy guarantee that enables us to quantify and analyze the privacy benefits of mixup augmentation.

Differential privacy, developed by Dwork et al. (2014), aims to prevent the leakage of potentially compromising information about individuals present in released data sets. By utilizing noise and randomness, differentially private data release mechanisms are provably robust to any auxiliary information available to an adversary.

Formally, let $\mathcal{M} : \mathcal{D} \to \mathcal{R}$ be a random mechanism, mapping from the space of datasets to a co-domain containing potential outputs of the mechanism. We consider a special case where $\mathcal{R}$ is another space of datasets, so that $\mathcal{M}$ outputs a synthetic dataset. We say two datasets $D, D' \in \mathcal{D}$ are adjacent if they differ by at most one element, that is D' has one fewer, one more, or one element different from D.

Then, $\mathcal{M}$ is (ϵ, δ)-differentially private if it satisfies the following inequality for any $U \subseteq \mathcal{R}$:

$$\mathbb{P}[\mathcal{M}(D) \in U] \leq e^{\epsilon}\mathbb{P}[\mathcal{M}(D') \in U] + \delta. \tag{1}$$

Intuitively, the inequality and symmetry in the definition of dataset adjacency tells us that the probability of getting any outcome from $\mathcal{M}$ does not strongly depend on the inclusion of any individual in the dataset. In other words, given any outcome of the mechanism, a strong privacy guarantee implies one cannot distinguish whether D or D' was used to produce it. This sort of indistinguishability condition is what grants protection from linkage attacks such as those explored by Narayanan & Shmatikov (2006). The quantity ϵ describes the extent to which the probabilities differ for *most* outcomes, and δ represents the probability of observing an outcome which *breaks* the ϵ guarantee.

In the case where differentially private datasets are used to train neural networks, such indistinguishability also assures poisoned data will not have a large effect on the trained model. Ma et al. (2019) formalize this intuition by proving a lower bound for the defensive capabilities of differentially private learners against poisoning attacks.

We define the threat model as taken from Ma et al. (2019): The attacker aims to direct the trained model $\mathcal{M}(D')$ to reach some attack target by modifying at most l elements of the clean dataset D to produce the poisoned dataset D'. We measure the distance of $\mathcal{M}(D')$ from the attack target using a cost function C, which takes trained models as an input and outputs an element of $\mathbb{R}$. The attack problem is then to minimize the expectation of the cost of $\mathcal{M}(D')$.

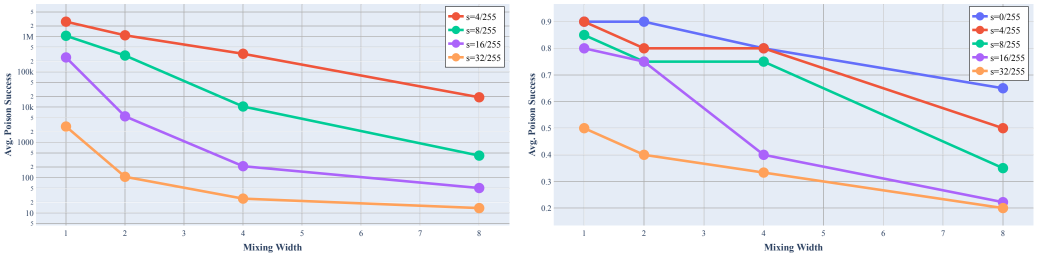
$$\min_{D'} J(D') := \mathbb{E}[C(\mathcal{M}(D'))] \tag{2}$$

Finally, we arrive at the theorem proven in Ma et al. (2019).

Theorem 1. *For an (ϵ, δ)-differentially private mechanism $\mathcal{M}$ and bounded cost function $|C| \leq B$, it follows that the attack cost $J(D')$ satisfies*

$$J(D') \geq \max\{e^{-l\epsilon}\left(J(D) + \frac{B\delta}{e^{\epsilon}-1}\right) - \frac{B\delta}{e^{\epsilon}-1}, 0\} \tag{3}$$

$$J(D') \geq \max\{e^{-l\epsilon}\left(J(D) + \frac{B\delta}{e^{\epsilon}-1}\right) + \frac{B\delta}{e^{\epsilon}-1}, -B\} \tag{4}$$



(a) Theoretical privacy guarantees

(b) Empirical defense against poisoning attacks

Figure 1: Theoretical and empirical mixup. Left: Privacy guarantee ϵ as a function of mixture width k, computed for each implemented Laplacian noise level s. We use values $n = T = 5 \times 10^4$, corresponding to the CIFAR-10 dataset. Right: Poisoning success for a strong adaptive gradient matching attack for several mixture widths and noise levels.

where the former bound holds for non-negative cost functions and the latter holds for non-positive cost functions.

Empirically, however, it is found that the defense offered by differential privacy mechanisms tends to be more effective than the theoretical limit. Likely, this is a result of differential privacy definitionally being a worst-case guarantee, and in practice the worst case is rarely observed.

We find that differential privacy achieved through the combination of k-way mixup and additive Laplacian noise is an example of such a defense. Because mixup augmentation concentrates training data near the center of the unit hypercube, less noise must be added to the mixed up data to render the noisy data indistinguishable from other points nearby in comparison to solely adding noise to the data points (Zhang et al., 2017). Additionally, mixup benefits from improved generalization due to its enforcement of linear interpolation between classes and has recently been shown to be robust to a variety of adversarial attacks, such as FGSM Zhang et al. (2020). We use a combinatorial approach to achieve a formal differential privacy guarantee for mixup with Laplacian noise, which in tandem with the result from Ma et al. (2019) gives us a direct theoretical protection from data poisoning.

2.1 A Theoretical Guarantee for DP-InstaHide

Let D be a dataset of size n and D' denote the same dataset with the point x_0 removed. Let d be the dimension of data points and assume the data lies in a set V of diameter one, i.e., $sup\{||D - D'||_1 : D, D' \in V\} \leq 1$. We sample a point of the form $z = \frac{1}{k}(x_1 + x_2 + \cdots + x_k) + \eta$, where the x_i are drawn at random from the relevant dataset P without replacement, and $\eta \sim Lap(\mathbf{0}, \sigma I)$ is the independent d-dimensional isotropic Laplacian additive noise vector with density function $\phi_\sigma(\eta) = \frac{1}{(2\sigma)^d} e^{\|\eta\|_1/\sigma}$. The random variable representing the outcome of the sampling is therefore a sum of random variables:

$$\mathcal{M}_P = \frac{1}{k} \sum_{i=1}^{k} X_i + N \tag{5}$$

Our differential privacy guarantee is stated below and proven in section A.2.

Theorem 2. *Assume the data set D has ℓ_1-norm radius less than 1, and that mixup groups of mixture width k are sampled without replacement. The DP-InstaHide method producing a data set of size T satisfies $(\epsilon, 0)$-differential privacy with*

$$\epsilon = T \max\{A, B\} \leq \frac{T}{k\sigma}$$

where

$$A = \log\left(1 - \frac{k}{n} + e^{\frac{1}{k\sigma}} \frac{k}{n}\right), \quad B = \log \frac{n}{n - k + ke^{-\frac{1}{k\sigma}}}.$$

Remark: A classical Laplacian mechanism for differentially private dataset release works by adding noise to each dataset vector separately and achieves privacy with $\epsilon = \frac{1}{\sigma}$. Theorem 2 recovers this bound in the case $k = 1$, however it also shows that k-way mixup enhances the privacy guarantee over the classical mechanism *by a factor of at least* k.

2.2 Defending with DP Augmentations in Practice

We investigate the practical implications of Theorem 2 in Figure 1, where we show the predicted theoretical privacy guarantees in Figure 1a and the direct practical application for defenses against data poisoning in Figure 1b. Figure 1b shows the average poison success for a strong, adaptive gradient matching attack against a ResNet-18 trained on CIFAR-10 (the setting considered in Geiping et al. (2020) with an improved adaptive attack). We find that the theoretical results predict the success of a defense by mixup with Laplacian noise surprisingly well.

As a result of Theorem 2, we investigate data augmentations with additional Laplacian noise, also in the setting of a gradient matching attack. Figure A.4 shows that the benefits of Laplacian noise which we only prove for mixup also extend empirically to variants of mixing data augmentations such as CutMix and MaxUp. In particular, combining MaxUp with Laplacian noise of sufficient strength ($s = 16/255$) completely shuts down the data poisoning attack via adaptive gradient matching, significantly improving upon numbers reached by MaxUp alone. Moreover, Figure A.2 and Table A1 show these data augmentations also exhibit a strong security performance trade-off compared to other defenses in the case of backdoor and gradient matching attacks.

References

Hojjat Aghakhani, Dongyu Meng, Yu-Xiang Wang, Christopher Kruegel, and Giovanni Vigna. Bullseye Polytope: A Scalable Clean-Label Poisoning Attack with Improved Transferability. *arXiv:2005.00191 [cs, stat]*, April 2020.

Nicholas Carlini, Samuel Deng, Sanjam Garg, Somesh Jha, Saeed Mahloujifar, Mohammad Mahmoody, Shuang Song, Abhradeep Thakurta, and Florian Tramer. An Attack on InstaHide: Is Private Learning Possible with Instance Encoding? *arXiv:2011.05315 [cs]*, November 2020.

Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian Molloy, and Biplav Srivastava. Detecting backdoor attacks on deep neural networks by activation clustering. *arXiv preprint arXiv:1811.03728*, 2018.

Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.

Jonas Geiping, Liam Fowl, W Ronny Huang, Wojciech Czaja, Gavin Taylor, Michael Moeller, and Tom Goldstein. Witches' brew: Industrial scale data poisoning via gradient matching. *arXiv preprint arXiv:2009.02276*, 2020.

Chengyue Gong, Tongzheng Ren, Mao Ye, and Qiang Liu. Maxup: A simple way to improve generalization of neural network training. *arXiv preprint arXiv:2002.09024*, 2020.

Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017.

Moritz Hardt and Kunal Talwar. On the geometry of differential privacy. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pp. 705–714, 2010.

W Ronny Huang, Jonas Geiping, Liam Fowl, Gavin Taylor, and Tom Goldstein. Metapoison: Practical general-purpose clean-label data poisoning. *arXiv preprint arXiv:2004.00225*, 2020a.

Yangsibo Huang, Zhao Song, Kai Li, and Sanjeev Arora. InstaHide: Instance-hiding Schemes for Private Distributed Learning. In *International Conference on Machine Learning*, pp. 4507–4518. PMLR, November 2020b.

Kangwook Lee, Hoon Kim, Kyungmin Lee, Changho Suh, and Kannan Ramchandran. Synthesizing differentially private datasets using random mixing. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pp. 542–546. IEEE, 2019.

Yuzhe Ma, Xiaojin Zhu, and Justin Hsu. Data poisoning against differentially-private learners: Attacks and defenses. *arXiv preprint arXiv:1903.09860*, 2019.

Arvind Narayanan and Vitaly Shmatikov. How to break anonymity of the netflix prize dataset. *arXiv preprint cs/0610105*, 2006.

Neehar Peri, Neal Gupta, W. Ronny Huang, Liam Fowl, Chen Zhu, Soheil Feizi, Tom Goldstein, and John P. Dickerson. Deep k-nn defense against clean-label data poisoning attacks, 2019.

Aniruddha Saha, Akshayvarun Subramanya, and Hamed Pirsiavash. Hidden trigger backdoor attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 11957–11965, 2020.

Avi Schwarzschild, Micah Goldblum, Arjun Gupta, John P Dickerson, and Tom Goldstein. Just how toxic is data poisoning? a unified benchmark for backdoor and data poisoning attacks. *arXiv preprint arXiv:2006.12557*, 2020.

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. *arXiv preprint arXiv:1804.00792*, 2018.

Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. *arXiv preprint arXiv:1811.00636*, 2018a.

Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. *arXiv preprint arXiv:1811.00636*, 2018b.

Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Clean-label backdoor attacks. *arXiv*, 2018.

Yu-Xiang Wang, Borja Balle, and Shiva Prasad Kasiviswanathan. Subsampled rényi differential privacy and analytical moments accountant. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1226–1235. PMLR, 2019.

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 6023–6032, 2019.

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

Linjun Zhang, Zhun Deng, Kenji Kawaguchi, Amirata Ghorbani, and James Zou. How does mixup help with robustness and generalization? *arXiv preprint arXiv:2010.04819*, 2020.

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*, pp. 7614–7623. PMLR, 2019.

A Appendix

Experimental details for the experiments shown in the main paper are contained in this document.

A.1 Data Augmentation as an Empirical Defense against Dataset Manipulation

We study the empirical effectiveness of data augmentations to prevent poisoning. We are mainly interested in data augmentations that mix data points; we consider the hypothesis that data poisoning attacks rely on the deleterious effects of a subset of modified samples, which can in turn be diluted and deactivated by mixing them with other, likely unmodified, samples.

One such augmentation is mixup, proposed in Zhang et al. (2017), which trains on samples $(x, y)_{i=1}^k$ mixed randomly in input space

$$\hat{x} = \sum_{i=1}^{k} \lambda_i x_i, \quad \hat{y} = \sum_{i=1}^{k} \lambda_i y_i, \tag{6}$$

to form the augmented sample $(\hat{x}, \hat{y})$. Though λ is traditionally drawn from a Dirichlet distribution parametrized by some chosen factor α, we will restrict to the case of equal weighting $\lambda = 1/k$ to aid in theoretical analysis. From here on, k is referred to as the mixture width.

CutOut (DeVries & Taylor, 2017), which blacks out a randomly generated patch from an image, can be combined with mixup to form CutMix (Yun et al., 2019), another type of mixing augmentation. Specifically, the idea is to paste a randomly selected patch from one image onto a second image, with labels computed by taking a weighted average of the original labels. The weights of the labels correspond to the relative area of each image in the final augmented data point.

MaxUp (Gong et al., 2020) can also be considered as a mixing data augmentation, which first generates augmented samples using various techniques and then selects the sample with the lowest associated loss value to train on. CutMix and mixup will be the central mixing augmentations that we consider in this work, which we contrast with MaxUp in select scenarios.

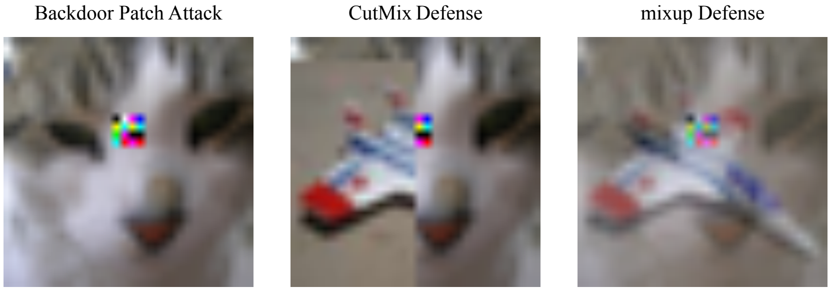


Figure A.1: "Cat" image from CIFAR-10 with a backdoor patch and the same image with CutMix and mixup augmentations.

Adding noise to input data is another augmentation method, which can be understood as a mixing augmentation that combines input data not with another image, but with a random sample from the input space, unrelated to the data distribution. This mechanism is also common in differential privacy (Hardt & Talwar, 2010). Since the exact original image is not apparent from its noised counterpart, adding noise decreases the sensitivity of the new data to the original dataset.

A.1.1 Backdoor Attacks

In contrast to recent targeted data poisoning attacks, *backdoor* attacks often involve inserting a simple preset trigger into training data to cause base images to be misclassified into the target class. For our experiments, we use small 4×4 randomly generated patches as triggers to poison the target class (See Figure A.1). To evaluate the baseline effectiveness of backdoor attacks, we poison a target class, train a ResNet-18 model on this poisoned data and use it to classify patched images from a victim test class. Only if a patched image from a victim class is labeled with the target class do we treat it

Table A1: Validation accuracy and poison success for a baseline model, models trained with mixup and CutMix augmentations (rows 2,3) and Spectral Signature Tran et al. (2018a) and Activation Clustering Chen et al. (2018) defenses (rows 4,5). The first two columns correspond to the case where 10% of one class is poisoned. The last two columns correspond to the case where all images of one class are poisoned (a scenario in which filter defenses are inapplicable as no unmodified images remain for this class). The results are averaged across 20 runs (with different pairs of target and victim classes).

	Clean Accuracy (10%)	Poison Success (10%)	Clean Accuracy (100%)	Poison Success (100%)
Baseline	94.3%	45.6%	85.0%	98.3%
CutMix	95.1%	**7.0%**	94.2%	**14.1%**
mixup	94.4%	23.9%	85.3%	99.8%
SS	92.3%	48.3%		
AC	89.4%	44.0%		

as a successfully poisoned example. Our results show that backdoor attacks achieve 98.3% poison success when 100% of images from the target class are poisoned and 45.6% poison success when only 10% of target images are patched (see Table A1). In addition, when 100% of training images from the target class are patched, clean test accuracy of the model drops by almost 10% since the model is unable to learn meaningful features of the target class.

We then compare the baseline model to models trained with the mixup and CutMix data augmentation techniques. We find that although mixup helps when only part of the target class is poisoned, it is not efficient as a defense against backdoor attacks when all images in the target class are patched. In contrast, CutMix is an extremely effective defense against backdoor attacks in both scenarios and it reduces poison success from 98.3% to 14.1% in the most aggressive setting. Finally, models trained on poisoned data with CutMix data augmentation have a clean test accuracy similar to the accuracy of models trained on clean data. Intuitively, CutMix often produces patch-free mixtures of the target class with other classes, hence the model does not solely rely on the patch to categorize images of this class.

We extend this analysis to two more complex attacks, clean-label backdoor attacks Turner et al. (2018), and hidden-Trigger backdoor attacks in Table A6.

For the patch attack, we insert patches of size 4×4 into CIFAR train images from target class and test images from victim class. The patches are generated using a Bernoulli distribution and are normalized using the mean and standard deviation of CIFAR training data. The patch location for each image is chosen at random. To evaluate the effectiveness of the backdoor attack and our proposed defenses, we train a ResNet-18 model on poisoned data with cross-entropy loss. The model is trained for 80 epochs using SGD optimizer with a momentum of 0.9, a weight decay of 5e-4 and learning rate of 0.1 which we reduce by a factor of 10 at epochs 30, 50 and 70. A batch size of 128 is used during training.

A.1.2 Targeted Data Poisoning

We further evaluate data augmentations as a defense against targeted data poisoning attacks. We analyze the effectiveness of CutMix and mixup as a defense against feature collision attacks in Table A4. Applying these data augmentations as a defense against Poison Frogs (Shafahi et al., 2018) (FC) is exceedingly successful, as the poisoned data is crafted independently there, making it simple to disturb by data augmentations. The poisons crafted via Convex Polytope (CP) (Zhu et al., 2019) however, are more robust to data augmentations, due to the polytope of poisoned data created around the target. Nonetheless, the effectiveness of CP is diminished more by data augmentations than by other defenses.

We then evaluate the success of data augmentations against Witches' Brew, the gradient matching attack of Geiping et al. (2020) in Table A2. Against this attack, we evaluate a wide range of data augmentations, as the attack is relatively robust to basic mixup data augmentations which mix only two images. However, using a stronger augmentation that mixes four images still leads to a strong defense in the non-adaptive setting (where the attacker is unaware of the defense). As this attack can be adapted to specific defenses, we also consider such a scenario. Against the adaptive attack, we found MaxUp to be most effective, evaluating the worst-case loss for every image in a minibatch over four samples of data augmentation drawn from cutout. To control for the effects of the CIFAR-10 dataset that we consider for most experiments, we also evaluate defenses against an attack on the ImageNet dataset in Table A3, finding that the described effects transfer to other datasets.

Table A2: Poison success rates (lower is better for the defender) for various data augmentations tested against the gradient matching attack of Geiping et al. (2020). All results are averaged over 20 trials. We report the success of both a non-adaptive and an adaptive attacker.

AUGMENTATION	NON-ADAPTIVE	ADAPTIVE
2-WAY MIXUP	45.00%	72.73%
CUTOUT	60.00%	81.25%
CUTMIX	75.00%	60.00%
4-WAY MIXUP	5.00%	55.00%
MAXUP-CUTOUT	5.26%	20.00%

Table A3: Success rate for selected data augmentation when tested against the gradient matching attack on the ImageNet dataset. All results are averaged over 10 trials.

AUGMENTATION	POISON SUCCESS
NONE	90%
2-WAY MIXUP	50.00%
4-WAY MIXUP	30.00%

A.1.3 COMPARISON TO OTHER DEFENSES

We show that our method outperforms filter defenses when evaluating backdoor attacks, such as in Table A1 and Table A6, as well as when evaluating targeted data poisoning attacks, as we show for Poison Frogs and Convex Polytope in Table A4 and for Witches' Brew in Table A3 and A5. We note that data augmentations do not require additional training compared to filter defenses in some settings and are consequently more computationally efficient.

In Figure A.2, we plot the average poison success against the validation error for adaptive gradient matching attacks. We find that data augmentations exhibit a stronger security performance trade-off compared to other defenses.

Table A4: Poison success rate for Poison Frogs (Shafahi et al., 2018) and Convex Polytope (Zhu et al., 2019) attacks when tested with baseline settings and when tested with mixup and CutMix. All results are averaged over 20 trials.

ATTACK	BASELINE	SS	AC	MIXUP	CUTMIX
FC	80%	70%	45%	**5%**	**5%**
CP	95%	90%	75%	70%	**50%**

We run our experiments for feature collision attacks in Table 4 by likewise using the framework of Schwarzschild et al. (2020), running the defense with the same settings as proposed there and following the constraints considered in this benchmark. For gradient matching we likewise implement a number of data augmentations as well as input noise into the framework of Geiping et al. (2020).

Table A5: Poison success rates (lower is better for the defender) for competing defenses when tested against the gradient matching attack compared to mixup. For DP-SGD, we consider a noise level of $n = 0.01$. All results are averaged over 20 trials.

Defense	Poison Success
Spectral Signatures	95.00%
deepKNN	90.00%
Spectral Signatures	95.00%
Activation Clustering	30.00%
DP-SGD	86.25%
4-way mixup	5.00%

Table A6: Success rate against backdoor attacks when tested with baseline settings and when tested with the mixup and CutMix. All results are averaged over 20 trials.

Attack	Baseline	SS	AC	mixup	CutMix
HTBD	60%	65%	55%	20%	**10%**
CLBD	65%	60%	45%	25%	**15%**

We run all gradient matching attacks within their proposed constraints, using a subset of 1% of the training data to be poisoned for gradient matching and an ℓ^∞ bound of 16/255. For all experiments concerning gradient matching we thus consider the same setup of a ResNet-18 trained on normalized CIFAR-10 with horizontal flips and random crops of size 4, trained by Nesterov SGD with 0.9 momentum and 5e-4 weight decay for 40 epochs for a batch size of 128. We drop the initial learning rate of 0.1 at epochs 14, 24 and 35 by a factor of 10. For the ImageNet experiments we consider the same hyperparameters for an ImageNet-sized ResNet-18, albeit for a smaller budget of 0.01% as in the original work.

Comparing to poison detection algorithms, we re-implement *spectral signatures* (Tran et al., 2018b), *deep K-NN* (Peri et al., 2019) and *Activation Clustering* (Chen et al., 2018) with hyperparameters as proposed in their original implementations. For differentially private SGD, we implement Gaussian gradient noise and gradient clipping to a factor of 1 on the mini-batch level (otherwise the ResNet-18

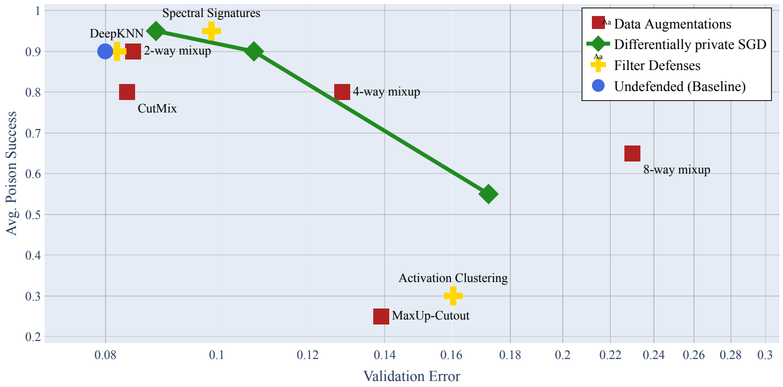


Figure A.2: Trade-off between average poison success and validation accuracy for various defenses against gradient matching (adaptive).

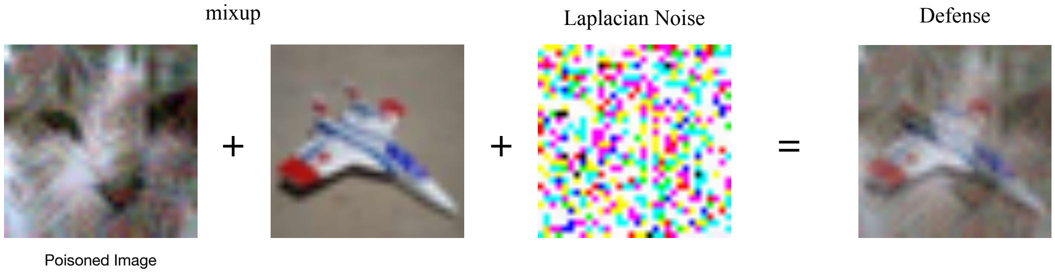


Figure A.3: Illustration of the DP-InstaHide defense on two CIFAR-10 images, the first of which has been poisoned with $\varepsilon = 16$. Mixup is used to average two images, and then Laplacian noise is added,

architecture we consider would be inapplicable due to batch normalizations), and vary the amount of gradient noise with values $(0.0001, 0.001, 0.01)$ to produce the curve in Fig. 2.

To implement data augmentation defenses we generally these data augmentations straightforward as proposed in their original implementations, also keeping components such as the late start of Maxup after 5 epochs described in Gong et al. (2020) and the randomized activation of CutMix described in Zhang et al. (2017).

A.2 A Theoretical Guarantee for DP-InstaHide

Let D be a dataset of size n and D' denote the same dataset with the point x_0 removed. Let d be the dimension of data points and assume the data lies in a set V of diameter one, i.e., $sup\{||D - D'||_1 : D, D' \in V\} \leq 1$. We sample a point of the form $z = \frac{1}{k}(x_1 + x_2 + \cdots + x_k) + \eta$, where the x_i are drawn at random from the relevant dataset P without replacement, and $\eta \sim Lap(\mathbf{0}, \sigma I)$ is the independent d-dimensional isotropic Laplacian additive noise vector with density function $\phi_\sigma(\eta) = \frac{1}{(2\sigma)^d} e^{\|\eta\|_1/\sigma}$. The random variable representing the outcome of the sampling is therefore a sum of random variables:

$$\mathcal{M}_P = \frac{1}{k} \sum_{i=1}^{k} X_i + N \tag{7}$$

We use p and q to denote the probability density functions of $\mathcal{M}_D$, and $\mathcal{M}_{D'}$ respectively.

Theorem. *Assume the data set D has ℓ_1-norm radius less than 1, and that mixup groups of mixture width k are sampled without replacement. The DP-InstaHide method producing a data set of size T satisfies $(\epsilon, 0)$-differential privacy with*

$$\epsilon = T \max\{A, B\} \leq \frac{T}{k\sigma}$$

where

$$A = \log\left(1 - \frac{k}{n} + e^{\frac{1}{k\sigma}} \frac{k}{n}\right), \; B = \log \frac{n}{n - k + ke^{-\frac{1}{k\sigma}}}.$$

Proof. To prove differential privacy, we must bound the ratio of $\mathbb{P}[\mathcal{M}_D \in U]$ to $\mathbb{P}[\mathcal{M}_{D'} \in U]$ from above and below, where $U \subseteq V$ is arbitrary and measurable. For a fixed sampling combination $x = (x_1, \ldots, x_k) \in D^k$, the density for observing $z = \frac{1}{k}\sum_{i=1}^{k} x_i + N$ is given by $\phi_\sigma\left(z - \sum_{i=1}^{k} x_i\right)$. Since there are $\binom{n}{k}$ possible values that x can take on, each of equal probability, we have

$$p(z) = \frac{k!(n-k)!}{n!} \sum_{x \in D^k} \phi_\sigma\left(z - \sum_{i=1}^{k} x_i\right).$$

Let's now write a similar expression for $q(z)$. We have

$$q(z) = \frac{k!(n-k-1)!}{(n-1)!} \sum_{x \in {D'}^k} \phi_\sigma\left(z - \sum_{i=1}^{k} x_i\right). \tag{8}$$

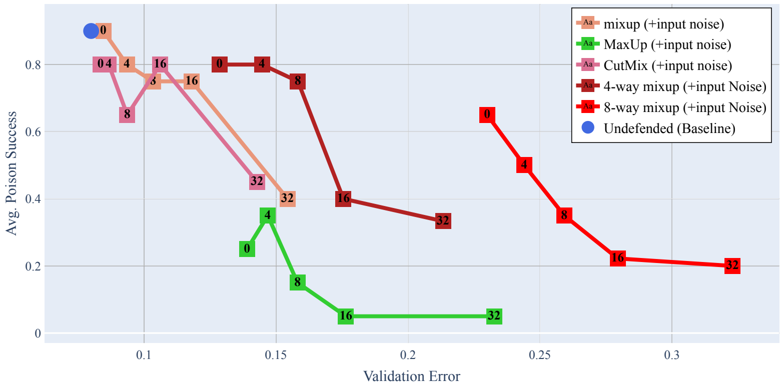


Figure A.4: Enhancing various data augmentations with Laplacian noise. We visualize the security-performance trade-off when enhancing the data augmentations considered in Sec. A.1 with Laplacian noise as predicted by Thm. 2. We visualize the development of these data augmentations when adding Laplacian noise with scales $(2/255, 4/255, 8/255, 16/255, 32/255)$.

Now, we write the decomposition $p(z) = p_0(z) + p_1(z)$, where $p_0(z)$ is the probability of the ensemble not containing x_0 times the conditional density for observing z given this scenario, and $p_1(z)$ is the probability of having x_0 in the ensemble times the conditional density for observing z given this scenario.

Then, we have

$$p_0(z) = \left(1 - \frac{k}{n}\right) q(z). \tag{9}$$

Now, consider $p_1(z)$. This can be written

$$p_1(z) = \frac{k}{n} \frac{(k-1)!(n-k-2)!}{(n-1)!} \sum_{x \in D'^{k-1}} \phi_\sigma \left(z - x_0 - \sum_{i=1}^{k-1} x_i \right). \tag{10}$$

In the equation above, $\frac{k}{n}$ represents the probability of drawing an ensemble x that contains x_0, and the remainder of the expression is the probability of forming $z - x_0$ using the remaining $k-1$ data points in the ensemble.

We can simplify equation equation 10 using a combinatorial trick. Rather than computing the sum over all tuples of size $k-1$, we compute the sum over all tuples of length k, but we discard the last entry of each tuple. We get

$$p_1(z) = \frac{k}{n} \frac{k!(n-k-1)!}{(n-1)!} \sum_{x \in D'^k} \phi_\sigma \left(z - x_0 - \sum_{i=1}^{k-1} x_i \right). \tag{11}$$

Now, from the definition of the Laplace density, we have that if $\|u - v\|_1 < \epsilon$ for any u, v then

$$e^{-\|u-v\|_1/\sigma} \phi_\sigma(v) \le \phi_\sigma(u) \le e^{\|u-v\|_1/\sigma} \phi_\sigma(v).$$

Let's apply this identity to equation 11 with $u = z - x_0 - \sum_{i=1}^{k-1} x_i$ and $v = z - \sum_{i=1}^{k} x_i$. We get

$$e^{-\frac{1}{k\sigma}} \frac{k}{n} q(z) \le p_1(z) \le e^{\frac{1}{k\sigma}} \frac{k}{n} q(z),$$

where we have used the fact that the dataset D has unit diameter to obtain $\|u - v\|_1 \le \frac{j}{k}$, and we used the definition equation 8 to simplify our expression.

Now, we add equation 9 to this equation. We get

$$\begin{aligned} \left(1 - \frac{k}{n} + e^{-\frac{1}{k\sigma}} \frac{k}{n}\right) q(z) &\le p(z) \\ &\le \left(1 - \frac{k}{n} + e^{\frac{1}{k\sigma}} \frac{k}{n}\right) q(z). \end{aligned}$$

From this, we arrive at the conclusion

$$\frac{p(z)}{q(z)} \leq \left(1 - \frac{k}{n} + e^{\frac{1}{k\sigma}} \frac{k}{n}\right) \leq e^{\frac{1}{k\sigma}},$$

and

$$\frac{q(z)}{p(z)} \leq \frac{n}{n - k + ke^{-\frac{1}{k\sigma}}} \leq e^{\frac{1}{k\sigma}}.$$

The left-most upper bound in the above equation is achieved by replacing k with n wherever k appears outside of an exponent. We get the final result by taking the log of these bounds and using the composibility property of differential privacy to account for the number T of points sampled.

□