Cluster Computing (2023) 26:349–366
https://doi.org/10.1007/s10586-022-03596-1

An empirical analysis of image augmentation against model inversion attack in federated learning

Seunghyeon Shin[1,2] · Mallika Boyapati[1] · Kun Suo[2] · Kyungtae Kang[3] · Junggab Son[1,2]

Received: 6 November 2021 / Revised: 5 February 2022 / Accepted: 4 April 2022 / Published online: 14 May 2022

Abstract
Federated Learning (FL) is a technology that facilitates a sophisticated way to train distributed data. As the FL does not expose sensitive data in the training process, it was considered privacy-safe deep learning. However, a few recent studies proved that it is possible to expose the hidden data by exploiting the shared models only. One common solution for the data exposure is differential privacy that adds noise to hinder such an attack, however, it inevitably involves a trade-off between privacy and utility. This paper demonstrates the effectiveness of image augmentation as an alternative defense strategy that has less impact of the trade-off. We conduct comprehensive experiments on the CIFAR-10 and CIFAR-100 datasets with 14 augmentations and 9 magnitudes. As a result, the best combination of augmentation and magnitude for each image class in the datasets was discovered. Also, our results show that a well-fitted augmentation strategy can outperform differential privacy.

Keywords Federated learning · Model inversion attack · Image augmentation · Defensive augmentation · Differential privacy

1 Introduction

Deep learning is being used in diverse areas including the medical field, engineering, fraud detection, and so on. As the purpose of deep learning is to learn about the huge amount of information to perform the task, collecting enough amount of data is an essential task to guarantee the performance of deep learning. However, collecting the data for training purposes has inherent risks as the data may contain sensitive information, so it may cause serious privacy infringement if not used properly or attacked. To deal with the given privacy issues, the Google AI team introduce a new training concept called Federated Learning, which enables participants to perform the learning process collaboratively without exposing their data that might include sensitive information [1]. In the Federated Learning process, a baseline model is distributed to participants, and the participants perform training with their data. The trained model parameter or the gradient is collected for the update of the given model when the training is completed, and the process repeats until the model returns satisfactory results. Federated learning is considered a relatively safe

✉ Junggab Son
json4@kennesaw.edu

Seunghyeon Shin
sshin9@students.kennesaw.edu

Mallika Boyapati
mboyapat@students.kennesaw.edu

Kun Suo
ksuo@kennesaw.edu

Kyungtae Kang
ktkang@hanyang.ac.kr

[1] Information and Intelligent Security (IIS) Laboratory, Kennesaw State University, 1100 South Marietta Pkwy, Marietta 30060, GA, USA

[2] Department of Computer Science, Kennesaw State University, 1100 South Marietta Pkwy, Marietta 30060, GA, USA

[3] Department of Computer Science and Engineering, Hanyang University, 55 Hanyangdeahak-ro, Ansan 15588, South Korea

method against privacy issues because it only collects the trained model parameters instead of data itself.

A few recent studies proved that stealing the hidden data from the trained model is not an impossible task. For example, Shokri et al. introduced a membership inference attack that figures out whether certain information is included in a training data or not [2], and Fredrikson et al. introduced a model inversion attack that extracts the target information from the model parameters [3]. The attackers have a significant advantage in stealing data especially when the training model is open to the public, such as when the developers utilize machine learning as a service (MaaS). Differential privacy is a widely used solution that protects sensitive data from attackers. The main idea of differential privacy is to protect sensitive data from attacks by perturbing the queries with randomly distributed noise, so the adversary cannot find if the target information is included in the dataset or not. However, a few issues exist regarding the deployment of differential privacy. As the deployment of differential privacy requires the usage of noise to defend intrusive queries from attackers, it inevitably involves the trade-off between utility and privacy caused by the noise. Also, the complicated background of differential privacy impedes the optimized implementation even for professionals.

To overcome the inherent issues of differential privacy, we introduce a new defense strategy against the proposed privacy attacks based on the augmentation of data. The original purpose of data augmentation is to facilitate the feature extraction process of the training model by increasing the amount of data through alterations or synthesis of new data, therefore the model could return better performance. The idea of this research is based on the observations that image data is often significantly modified when some types of augmentations are applied to it, so it might be used like the randomized noise deployed in differential privacy-based defense strategies that protect the sensitive data.

This paper provides meaningful results by performing comprehensive experiments, which demonstrates the effectiveness of the augmentation strategies as a potential way to defend the attacks while preserving enough utility. Our experiments utilize CIFAR-10 and CIFAR-100 datasets with 14 different augmentations and the magnitude range of 1 to 9, and as a result, we could find the optimized augmentation strategies for each label of the dataset that outperforms the result of differential privacy-based defense strategy. In the case of the CIFAR-10, for example, the solarize augmentation that inverts the pixels above threshold shows 16% and 11.6% better performance in model accuracy and attack accuracy with magnitude 7, which inverts approximately 78% of pixels from an image. The posterize augmentation presents approximately 4.47% and 12.42% better model and attack accuracy when it reduces 7-bit from each RGB channel. In the case of the CIFAR-100, we separately sampled 10 labels from the dataset that returns notable performance and the other 10 labels that returns limited performance compared to differential privacy-based defense strategies. The difference between the accuracy of conventional differential privacy and our augmentation-based defense strategy is provided as a form of the advantage score. Finally, we discuss future research that finds the optimized augmentation strategies for the given image type through deep reinforcement learning. In addition to the results, how augmentations applied to image data modifies, and how it successfully defends the leakage of training data are discussed.

2 Literature review

2.1 Model inversion attack

Fredrikson et al. proposed the first model inversion attack method against neural network in 2014 [3]. The research showed that the adversary can infer the genotype of the victim from a linear regression model with black-box access and some non-sensitive attributes. Fredrikson et al. published extended research of model inversion attack that recovers the image data from facial recognition system, but it could not reconstruct recognizable object from the setting [4]. Hidano et al. introduced an enhanced model of Fredrikson's research [3] that does not require non-sensitive attributes by injecting the malicious data that adversary possesses to modify the target model [5]. In 2019, Zhu et al. showed that private data can be leaked from shared gradients by minimizing the difference between original gradient and dummy gradient [6]. Zhao et al. made continuous research that discovers the ground-truth image with improved optimization performance and less number of iterations [7]. However, the attacks proposed by Zhu et al. and Zhao et al. were only tested on a shallow neural network model and could not retrieve the big image data. The continuous research in model inversion attack method with notable reconstruction quality was introduced by Geiping et al. [8]. The research showed that neural networks can be attacked regardless its depth or image size. The authors also showed that they can perform multi-image reconstruction from model parameters at the same time.

2.2 Differential privacy

In previous model inversion research, many authors mentioned that differential privacy could be a solution for diverse attacks against privacy including model inversion attack [3, 9, 10]. The core concept of differential privacy is based on the usage of Laplacian and Gaussian noises based on l_1 norm and l_2 norm, which were introduced by Dwork et al. and mcsherry et al. [11, 12]. The recent researches adopted the idea of differential privacy for secure deep learning [13–16]. Shokri and Shmatikov introduced distributed selective stochastic gradient descent (DSSGD) that injects Laplacian noise into the optimization process for collaborative deep learning [13]. Abadi et al. proposed an improved strategy that utilizes Gaussian noise and moments account called differentially private stochastic gradient descent (DPSGD) to control the amount of injected noise, therefore enabling tracking the amount of privacy spent [14]. Phan et al. has increased the efficiency of training throughout the adaptive controlling of noise amount in the optimization process based on the importance of features [15]. Mironov proposed an enhanced differential privacy concept called Rényi Differential Privacy (RDP) based on Rényi divergence of order α, which measures the divergence between two adjacent datasets [16]. Truex et al. showed that differential privacy can be used in federated learning by adding a local differential privacy module that guarantees the privacy of sensitive data before sending parameters to a centralized server [17]. Girgis et al. proposed a stochastic gradient descent algorithm that inputs the sampled clients and the data points of chosen clients to the shuffler, therefore guaranteeing privacy by hiding which clients were chosen in a federated environment [18].

2.3 Data augmentation

The studies in data augmentation have been made with the development of image vision by exploring various methods, including but not limited to traditional augmentations, Generative Adversarial Network (GAN), reinforcement learning, and automated machine learning. A few innovative augmentation strategies were made, such as Cutout [19], and Random Erasing [20], which arbitrarily determines the area of an image to be masked out. In addition to that, Wu et al. introduced multiple color augmentation strategies that adjust color pixels to diversify input features [21]. Unlike the traditional augmentation strategies that modify the given input, Generative Adversarial Networks (GAN) creates the whole new data that is similar to the original image through the synthesis of given data. GAN-based augmentations can be utilized in classification [22, 23], privacy de-identification [24], synthesis of high resolution image [25, 26] and so on. A few recent researches adopted automated machine learning to find optimized augmentations based on reinforcement learning [27], Bayesian optimization [28], differentiation of policy search [29], and grid search [30].

3 Preliminaries

3.1 Federated learning

The purpose of Federated Learning is to facilitate the usage of data stored in distributed data centers. The idea of Federated Learning introduced by Mcmahan *et* al. does not require the data sharing between the centralized server and participants, but enables collaborative learning as a federation [1]. The formal description of Federated Learning is as follows: The centralized server distributes the model $W_1^k, W_2^k, \ldots, W_n^k$ to n local clients, and the clients train the received model W^k with their dataset. Then the local updates $U := W_i^k - W^k$ made by clients are aggregated for the averaging process. Using the gradient descent, the globally updated model $W^{k+1} = W^k - \eta U'$ is made by the server, where η denote as learning rate and $U^{'} = \frac{1}{n}\sum U^i$. Then $W^{'}$ is distributed to clients repeatedly for continuous updates. Figure 1 visualizes the overall process of federated learning.

3.2 Differentially private stochastic gradient descent

The concepts of differential privacy proposed by Dwork et al. [31] is as follows: given two datasets that has at most one different element denoted by $\Delta(D_1, D_2) \leq 1$, a randomized computation F gives (ε, δ)-privacy if

$$Pr[F(D_1) \in S] \leq e^{\varepsilon} \cdot Pr[F(D_2) \in S] + \delta, \tag{1}$$

where F means query and noise added to the query, S means all probable output of F, ε means the maximum distance between the same queries on D_1, D_2 that means privacy loss, and finally, δ means the probability of accidental information leakage. The additive noise for differential privacy is generated based on the Laplace mechanism or Gaussian mechanism. In deep learning, adopting the Gaussian mechanism has a few advantages

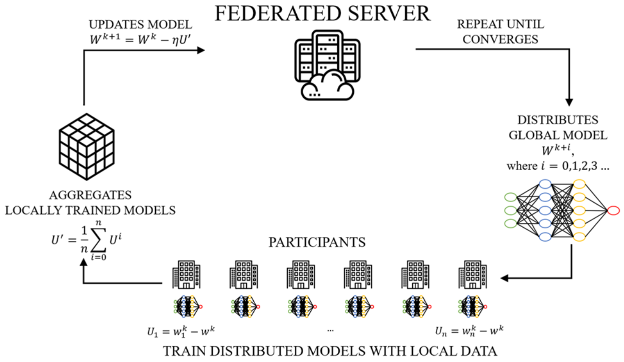

FEDERATED SERVER
UPDATES MODEL
$W^{k+1} = W^k - \eta U'$
REPEAT UNTIL CONVERGES
DISTRIBUTES GLOBAL MODEL
W^{k+i}, where $i = 0,1,2,3$ ...
AGGREGATES LOCALLY TRAINED MODELS
$U' = \frac{1}{n}\sum_{i=0}^{n} U^i$
PARTICIPANTS
$U_1 = w_1^k - w^k$
...
$U_n = w_n^k - w^k$
TRAIN DISTRIBUTED MODELS WITH LOCAL DATA

Fig. 1 Federated learning

over the Laplace mechanism: that is, it allows using either $l1$ sensitivity or $l2$ sensitivity depending on its purpose while the Laplace mechanism only allows the usage of $l1$ sensitivity, therefore it guarantees flexibility in adopting differential privacy. Also, it requires less amount of noise and a privacy budget when the Gaussian mechanism follows $l2$ sensitivity, which is significantly lower than $l1$ sensitivity followed by the Laplace mechanism. As the amount of noise significantly impacts the performance of a neural network that consists of a series of weights, [14] adopted Gaussian mechanism for differentially private deep learning in their study, which is denoted as follows:

given optimizer as Stochastic Gradient Descent, a gradient $g_t(x_i) = \nabla_\theta \mathcal{L}(\theta, x_i)$ from randomly selected small batches $\{x_1, x_2, ..., x_n\}$ is computed, and each gradient is clipped with clipping threshold C, which is denoted by $g_t(x_i)/max(1, \frac{\|g_t(x_i)\|}{C})$. Then the Gaussian noise $\mathcal{N}\ (0, \sigma^2)$ is injected, where $\mathcal{N}\ (0, \sigma^2)$ means Gaussian distribution with variance σ^2.

3.3 Exploitation of model for data reconstruction

The purpose of the model inversion attack is to extract the training data that is not open to the public. The proposed attack model for data reconstruction introduced in [8] steals significant information by exploiting model parameters shared in federated learning. The proposition is as follows: The participants are given the initial model parameter θ^k from a centralized server and train it with their data x_i and label y_i, then send back the gradient $\nabla \mathcal{L}(x_i, y_i)$ to have the server update the model parameter to θ^{k+1}. The adversary can extract the data from $\nabla \mathcal{L}(x_i, y_i)$ as the angle between two data points in gradient descent steps provides information that changes prediction. Previous researches [6, 7] minimizes the difference between the gradients from dummy data (x_i', y_i) and real data (x_i, y_i) by computing Euclidean distance:

$$||\nabla \mathcal{L}(x_i', y_i) - \nabla \mathcal{L}(x_i, y_i)||^2. \tag{2}$$

However, due to its inefficiency of computation and initialization issue against practical architectures used in recent studies, Geiping et al. minimizes gradient difference through cosine similarity that computes the similarity between gradient vectors [8]. The idea is denoted as follows:

$$\frac{\nabla \mathcal{L}(x_i, y_i) \cdot \nabla \mathcal{L}(x_i', y_i)}{max(||\nabla \mathcal{L}(x_i, y_i)|| \cdot ||\nabla \mathcal{L}(x_i', y_i)||, \varepsilon)}, \tag{3}$$

where ε refers to a small value that prevents division by zero. As the gradient of the image always includes significant information that can be extracted through cosine similarity, the adversary can always extract the training data even from pre-trained models. Figure 2 visualizes how the gap between gradient vectors is minimized.

3.4 Data augmentation

The purpose of data augmentation is to facilitate the extraction of input features by increasing the amount of training data $D = (x_1, x_2, ..., x_t)$ throughout the modification or the synthesis of data. As the amount of data increases from D to $D'' = D + D'$, where $D' = (x_1', x_2', ..., x_t')$, the training model can have better regularization performance, so overfitting issue can be mitigated. In general, augmentation applied to training data generally consists of geometric transformation and color space augmentation [32]. The geometric transformation includes various types of affine transformations that adjust the geometric location of an image while preserving the colinearity after transformation, and color space augmentation converts the RGB value of image pixels variously to remove biases in training data. Depending on the level of modification, applying data augmentation can significantly expand the size of training data with different features.

4 Materials and methods

4.1 System settings

In this section, we define the attack model introduced in [8] to compare the performance of differential privacy-based and augmentation-based defense strategies. The proposed attack model exploits the model parameters of the neural network to reconstruct the hidden training data. Unlike the attack models with limited performance introduced in previous research [6, 7], the proposed attack model works well for a realistic environment that utilizes a deep neural

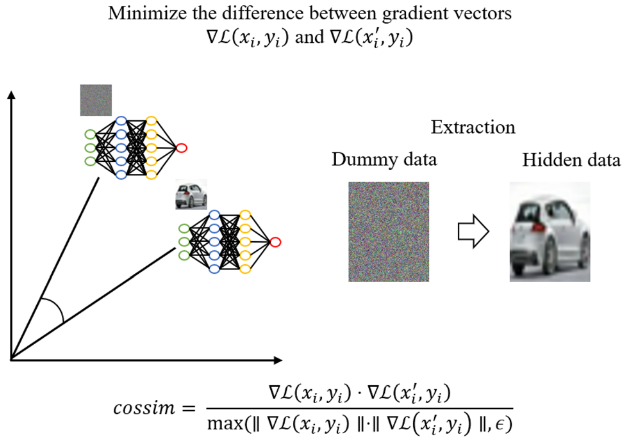

Minimize the difference between gradient vectors
$\nabla \mathcal{L}(x_i, y_i)$ and $\nabla \mathcal{L}(x_i', y_i)$
Extraction
Dummy data
Hidden data
$cossim = \frac{\nabla \mathcal{L}(x_i, y_i) \cdot \nabla \mathcal{L}(x_i', y_i)}{\max(\| \nabla \mathcal{L}(x_i, y_i) \| \cdot \| \nabla \mathcal{L}(x_i', y_i) \|, \epsilon)}$

Fig. 2 Model Inversion Attack

network and pretrained model. We conducted reconstruction experiments with Adam optimizer, learning rate 0.1, and set the maximum iteration to 4,000. As the number of model parameters affects the reconstruction time and quality, the model chosen for this experiment is the VGG-11 without the fully connected layers that balanced reconstruction time and quality. The customed VGG-11 model has approximately 9.23M parameters. In the reconstruction process, we reconstruct one image in each run to guarantee the highest quality of reconstructed images. Two datasets are used in this experiment: CIFAR-10 and CIFAR-100 where the number indicates the number of labels included in the dataset. Both datasets include 60,000 32×32 sized images equally separated in each label. In this experiment, we targeted reconstructing 30% of randomly sampled test data from each label to present reliable classification results. Two different types of accuracy called model accuracy and attack accuracy will be provided in this paper to compare the performance of conventional differential privacy-based defense strategy and our augmentation-based defense strategy. Model accuracy is the accuracy of the data before reconstruction and Attack accuracy is the accuracy of the reconstructed data. A total of 8 RTX-2080 GPUs was used to reconstruct 780,000 images, 390,000 for CIFAR-10 and CIFAR-100 each.

4.2 Differential privacy settings

Based on the idea of differential privacy that utilizes noise to perturb adversarial queries, we implemented a DP-SGD optimizer that stochastically scatters noise during the optimization process to compare its performance as a conventional defense strategy. The followings are the default hyperparameter settings: first, the clipping threshold that bounds the maximum gradient norm is set to 1, denoted by $C = 1$. We determined to utilize the Gaussian noise over Laplacian noise and the amount of Gaussian noise is controlled by its standard deviation σ. Three different amount of Gaussian noise is used in this experiment. For CIFAR-10, we utilized $\sigma =$ 0.1, 0.5, and 1.0. For CIFAR-100, we reduced the amount of noise to $\sigma =$ 0.1, 0.2, and 0.3 due to the sensitivity towards the amount of noise. Two famous open-source differential privacy libraries named PyVacy [33] and Opacus [34] were referenced for correct implementation. As we plan to defend against a reconstruction attack that steals training data, we prepared six pre-trained models trained with DP-SGD optimizer that injects different amounts of noise. The VGG-11 model introduced in system settings was trained 200 epochs each to create pre-trained models. Learning rate and δ was set to $1e^{-3}$ and $1e^{-5}$ each.

4.3 Augmentation settings

The augmentations tested in this paper include 14 different augmentations introduced in [27–30], which includes diverse color space augmentation and affine transformations. Each augmentation introduced below have 9 augmentation level from 1 to 9, which determines the level of modification from low to high. Figure 3 presents how the introduced augmentations modify the input data.

- Autocontrast - maximize the contrast of an image.
- Brightness - randomly adjusts the brightness of an image based on magnitude.
- Color - adjusts the balance of color of an image.
- Contrast - adjusts the contrast of an image based on magnitude.
- Equalize - adjusts the image histogram to be equalized. Histogram of an image refers to the distribution of pixels in a digital image.
- Invert - invert all pixels of an image.

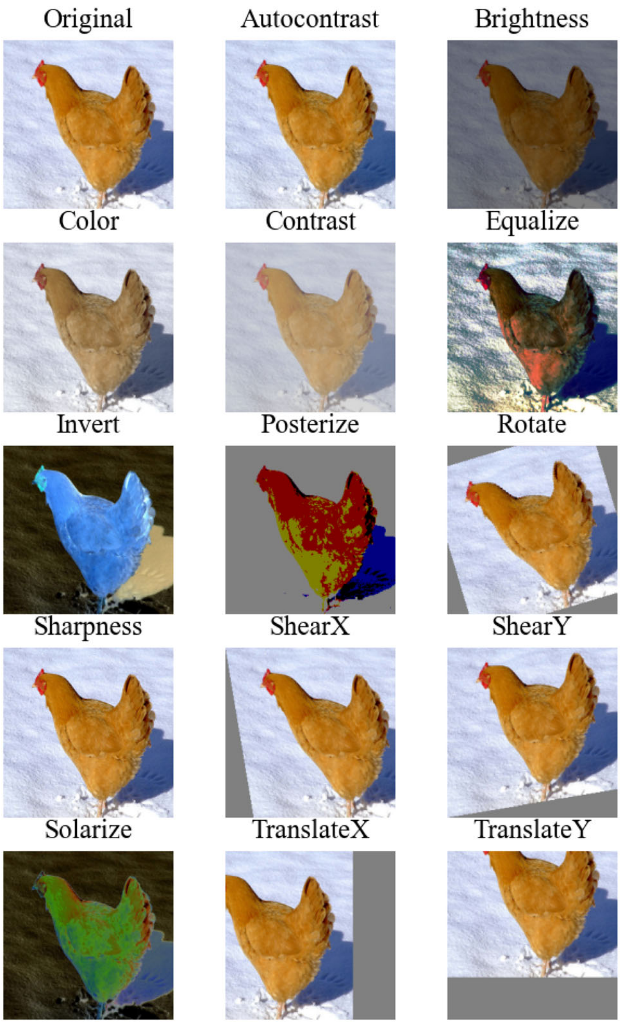


Fig. 3 Visualized augmentation strategies

Table 1 CIFAR-10 accuracy table for DP-SGD

	Original		$\sigma = 0.1$		$\sigma = 0.5$		$\sigma = 1.0$	
Accuracy(%)	MA	AA	MA	AA	MA	AA	MA	AA
Airplane	87.80	71.13	71.12	63.07	41.33	34.07	22.11	18.87
Automobile	93.73	77.20	83.53	76.33	63.80	56.73	48.80	40.40
Bird	74.20	69.73	54.30	49.80	32.63	28.47	18.00	13.67
Cat	62.94	55.60	40.21	36.87	38.51	34.53	33.54	30.47
Deer	82.00	82.80	58.92	54.07	36.17	33.47	28.40	24.27
Dog	66.80	65.07	54.70	51.93	46.86	42.80	37.73	29.40
Frog	84.27	82.13	84.65	81.87	69.91	68.87	52.67	47.93
Horse	84.80	83.07	74.13	70.87	46.25	44.80	35.40	29.87
Ship	92.94	85.60	79.48	75.93	47.72	41.73	30.33	25.60
Truck	86.67	83.53	86.65	80.73	81.94	78.73	73.13	67.27
Average	81.62	75.59	68.77	64.15	50.51	46.42	38.01	32.78

Table 2 CIFAR-100 accuracy table with DP-SGD optimizer

	Original		$\sigma = 0.1$		$\sigma = 0.2$		$\sigma = 0.3$	
Accuracy(%)	MA	AA	MA	AA	MA	AA	MA	AA
Average	62.43	44.60	39.37	24.07	34.42	12.3	29.7	5.33

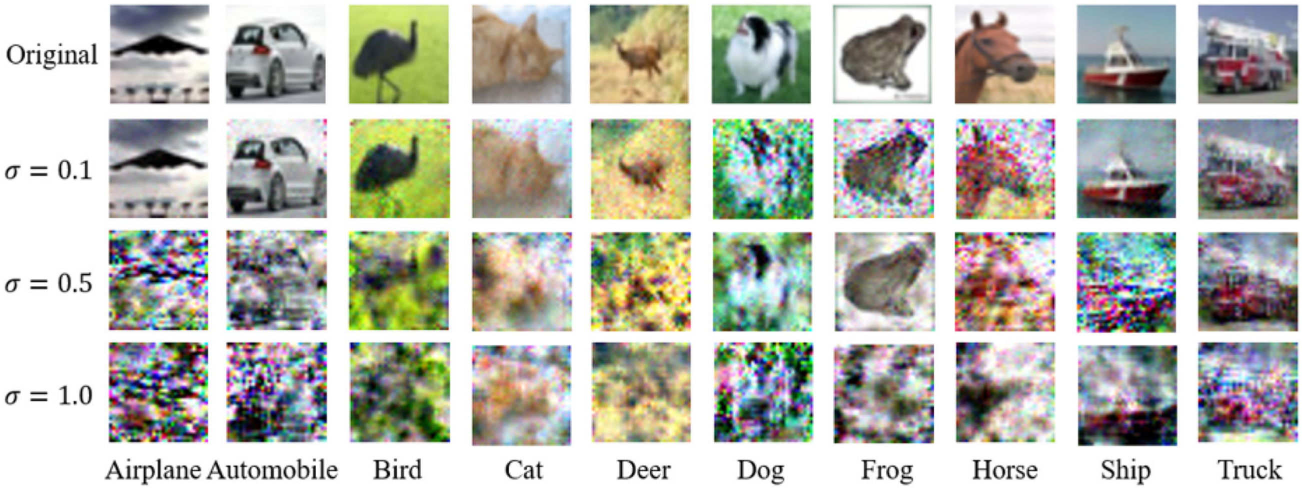


Fig. 4 Reconstructed CIFAR-10 images with Gaussian noise

- Posterize - Reduce the bits from each RGB color channel.
- Rotate - rotate an image.
- Sharpness - adjust the blurriness of an image. Low magnitude returns a sharper image.
- ShearX - shear an image vertically based on magnitude.
- ShearY - shear an image horizontally based on magnitude.
- Solarize - invert pixels of an image above magnitude.
- TranslateX - move an image along the X-axis.
- TranslateY - move an image along the Y-axis.

Due to the characteristics of an augmentation, three augmentations (autocontrast, equalize, invert) return the same images regardless of magnitude. Also, the reduction range of bits for posterize augmentation was defined from 0 to 4 bits in [27–30]. However, we extended the reduction range from 0 to 7 to maximize the effect of augmentation. Note that solarize inverts the input pixel above the threshold, so it returns the equivalent images with invert augmentation when the magnitude is set to 9 as the threshold is set to 0.

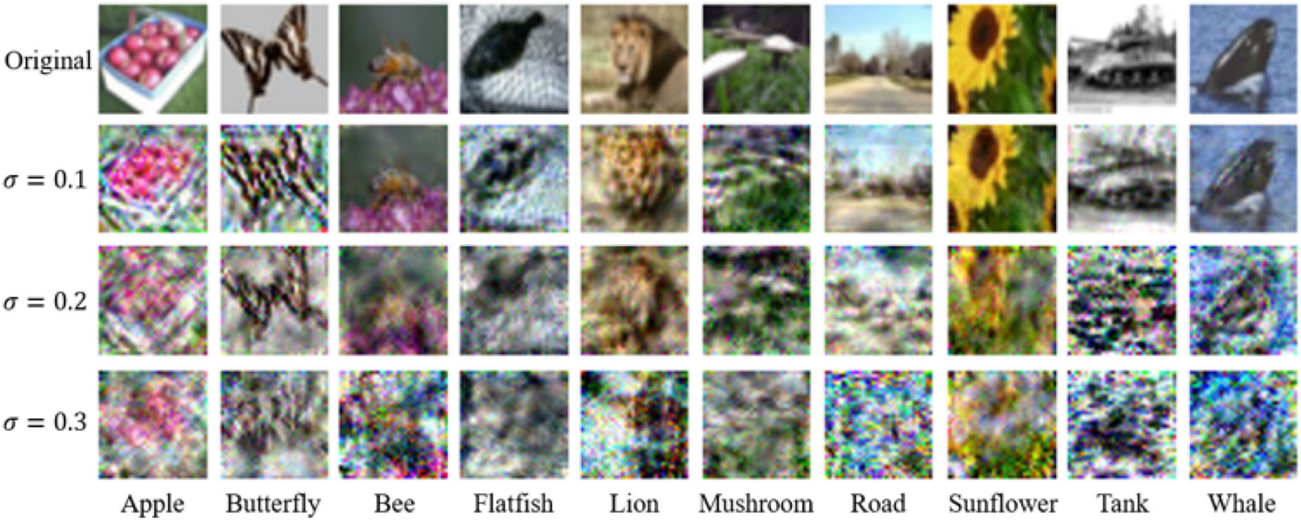


Fig. 5 Reconstructed CIFAR-100 images with Gaussian noise (10 samples)

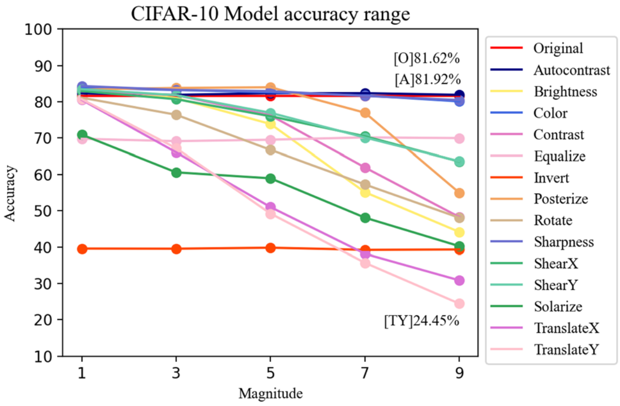


Fig. 6 Model accuracy range for CIFAR-10 augmentations

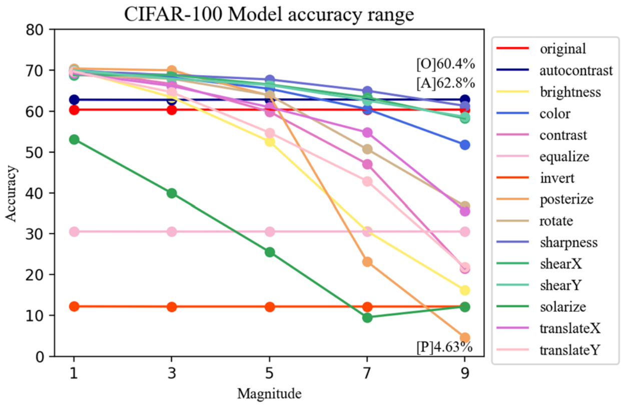


Fig. 8 Model accuracy range for CIFAR-100 augmentations

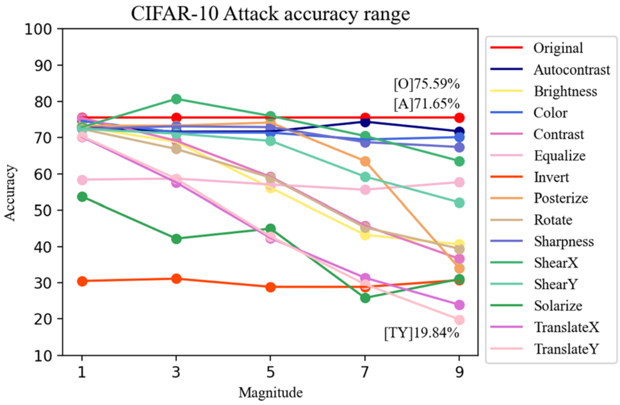


Fig. 7 Attack accuracy range for CIFAR-10 augmentations

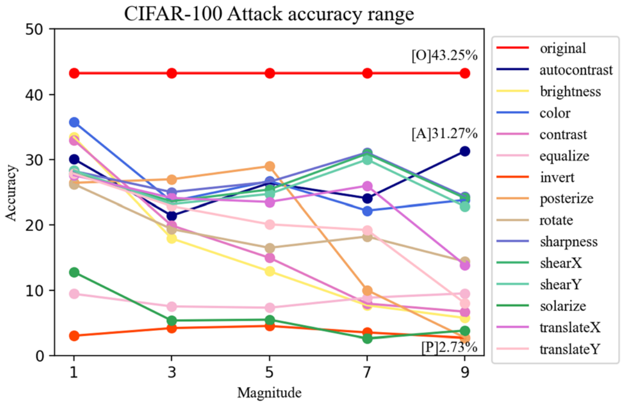


Fig. 9 Attack accuracy range for CIFAR-100 augmentations

Table 3 Optimized augmentations for CIFAR-10

Labels/Accuracy	1st highest	2nd highest	3rd highest
Airplane	Solarize (M7)	Equalize (M4)	Solarize (M1)
Automobile	Solarize (M7)	Solarize (M6)	Solarize (M3)
Bird	Contrast (M7)	Posterize (M8)	Posterize (M9)
Cat	Posterize (M9)	Contrast (M8)	Solarize (M4)
Deer	Brightness (M9)	TranslateX (M5)	TranslateY (M5)
Dog	Equalize (M7)	Solarize (M7)	Posterize (M9)
Frog	Contrast (M6)	Brightness (M5)	Solarize (M5)
Horse	Posterize (M9)	Posterize (M8)	Solarize (M6)
Ship	Solarize (M6)	Rotate (M9)	Posterize (M8)
Truck	Solarize (M3)	Brightness (M6)	Rotate (M4)

Table 4 Sampled augmentations for CIFAR-100

Labels	Efficient Augmentations	Labels	Inefficient Augmentations
Aquarium fish	Solarize (M3)	Chair	TranslateX (M2)
Beaver	Solarize (M2)	Dinosaur	Solarize (M6)
Beetle	Equalize (M5)	Fox	ShearY (M5)
Boy	Equalize (M9)	Hamster	Contrast (M5)
Can	Posterize (M9)	Keyboard	Sharpness (M7)
Lamp	Posterize (M9)	Mountain	Contrast (M3)
Motorcycle	Solarize (M2)	Racoon	Equalize (M6)
Sea	Brightness (M9)	Shark	ShearY (M7)
Tank	Posterize (M7)	Streetcar	Color (M7)
Worm	Contrast (M8)	Television	Invert (M9)

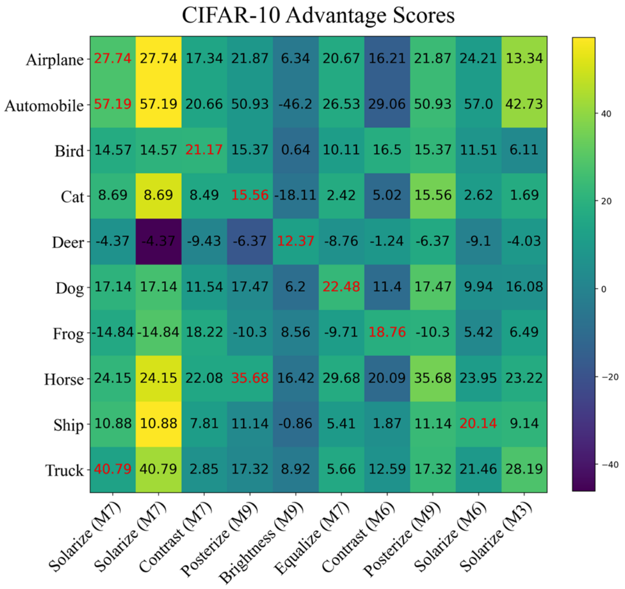


Fig. 10 CIFAR-10 advantage score matrix

5 Experimental results

Approximately 780,000 images (390,000 from each of CIFAR-10 and CIFAR-100) were reconstructed with various augmentations and differential privacy settings throughout the experiment. Firstly, we conducted reconstruction with the pre-trained models that were trained with DP-SGD optimizer with different levels of noise. Tables 1 and 2 below show differential privacy-based accuracy for CIFAR-10 and CIFAR-100. Note that model accuracy is the accuracy of original data with noise, and attack accuracy is the accuracy of reconstructed data with DP-SGD optimizer. ResNet-50 and VGG-11 without fully connected layers described previously were used for measuring accuracy and making pretrained models, respectively. As mentioned previously, DP-based defense inevitably involves utility loss as the noise affects the optimization performance in the classification process. In addition to that, DP-based defense still leaks some amount of training data in the reconstruction process even when enough amount of noise is injected as the noise is scattered randomly in the optimization process. The reconstruction time of each image is approximately 4 minutes for differential privacy settings, and 2 minutes for augmented images, on average.

Figures 4 and 5 show the inherent issues in differential privacy-based defense strategy. The frog image in Fig. 4 and the butterfly image in Fig. 5 are clearly recognizable regardless of the injected noise, and adding more noise

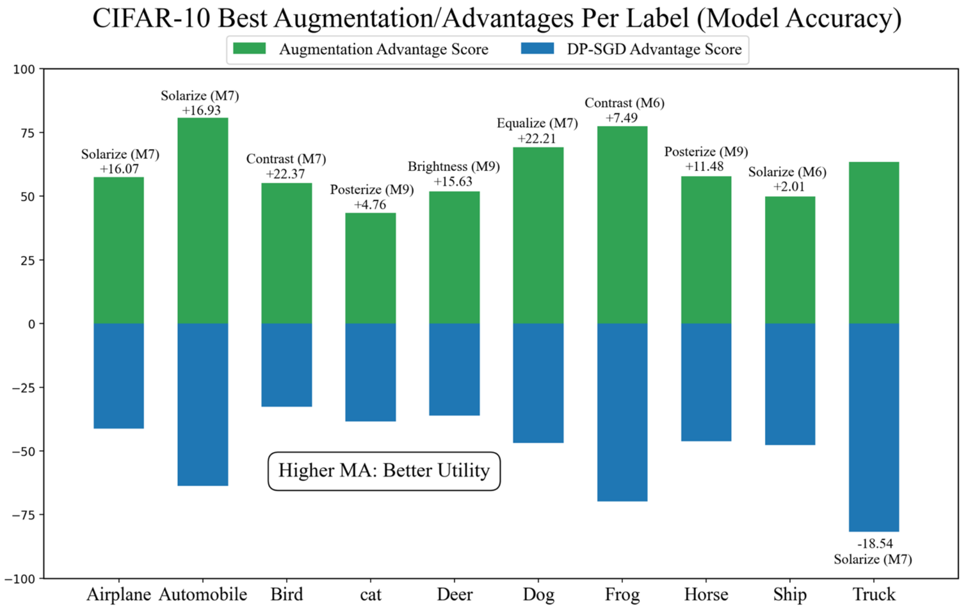


Fig. 11 CIFAR-10 model accuracy advantages

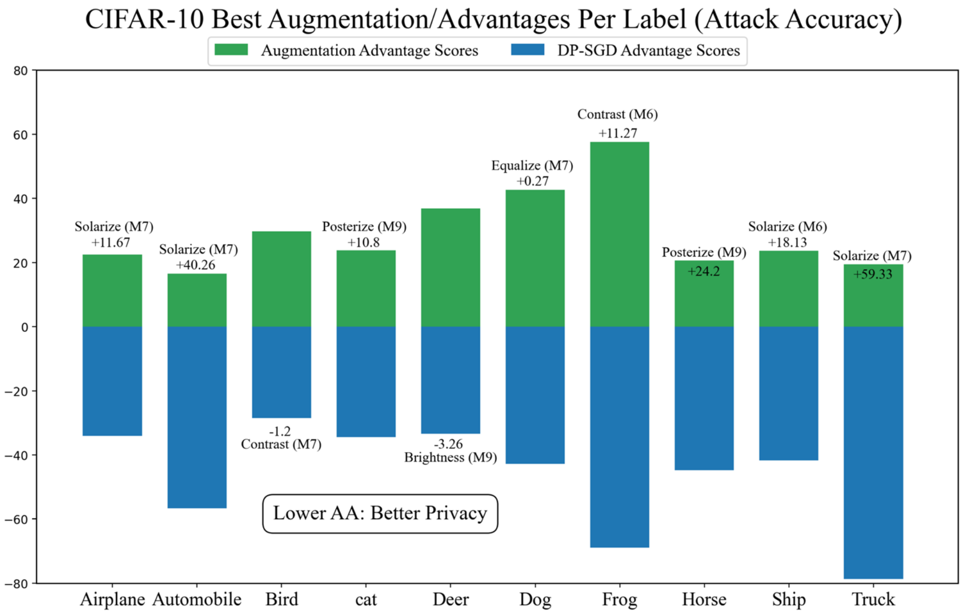


Fig. 12 CIFAR-10 attack accuracy advantages

Table 5 Advantage scores of best CIFAR-10 augmentations per label

Labels	Airplane	Automobile	Bird	Cat	Deer
Augmentation	Solarize (M7)	Solarize (M7)	Contrast (M7)	Posterize (M9)	Brightness (M9)
Advantage	27.74	57.19	21.17	15.56	12.37
Labels	Dog	Frog	Horse	Ship	Truck
Augmentation	Equalize (M7)	Contrast (M6)	Posterize (M9)	Solarize (M6)	Solarize (M7)
Advantage	22.48	18.76	35.68	20.14	40.79

significantly hurts the utility. We believe an augmentation-based defense strategy could be an enhanced solution for this problem depending on the type of augmentations and magnitude. In Figs. 6, 7, 8, and 9, linear graphs that show the accuracy change of CIFAR-10 and CIFAR-100 based on magnitude are provided. The provided figures show that augmentations applied to a dataset distort the given image

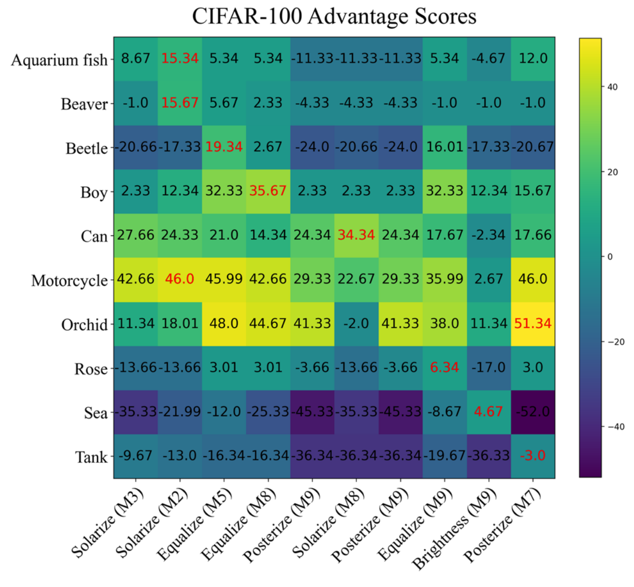


Fig. 13 CIFAR-100 advantage score matrix

data directly, therefore the accuracy decreases as the magnitude increases.

An advantage score is used to describe the efficiency of augmentation strategies in this paper, which is defined as follows:

$$\text{Advantage score} = (MA_{DP} - MA_{Aug}) + (AA_{DP} - AA_{Aug}), \quad (4)$$

where MA denotes a model accuracy, AA denotes an attack accuracy, DP is differential privacy with $\sigma = 0.5$ and $\sigma = 0.2$ respectively for CIFAR-10 and CIFAR-100, and Aug denotes an augmentation scheme. That is augmentations that return higher model accuracy and lower attack accuracy return a high advantage score. Tables 3 and 4 are the augmentation strategies that we have found based on the advantage score. The chosen augmentation with magnitude has lower attack accuracy and higher model accuracy than the DP-based defense strategy, which means it should successfully defend reconstruction attack while preserving a certain level of utility even when augmentation significantly distorts the original data. From the results above, we can see that color space augmentations have better efficiency compared to geometric transformation in the majority of cases. In Table 3, Geometric transformations such as TranslateX and TranslateY with magnitude 5 are only ranked in 2nd and 3rd best places for deer label, and rotate with magnitude 9 is ranked in 2nd for ship label. In Table 4, the inefficient augmentation columns returned notable advantage scores, however, they leaked a lot of information from the reconstructed data. The problem of geometric transformation is that it still leaks information from the area that is not augmented. As the geometric augmentations only affect a certain area of image data, it cannot be considered privacy-preserving even if it shows notable accuracy.

In Fig. 10, we provide the advantage score of CIFAR-10 and the best augmentations. A few labels return the best performance in model and attack accuracy when a certain type of augmentations are applied. For example, the airplane label in CIFAR-10 works best when solarize augmentation with magnitude 7 is applied, which shows 16.07% higher model accuracy and approximately 11.67% lower attack accuracy. In the case of the truck label located at the end of the graphs, it has slightly lower model accuracy, however, it shows significantly lower attack accuracy compared to DP-SGD when Solarize

Fig. 14 CIFAR-100 model accuracy advantages

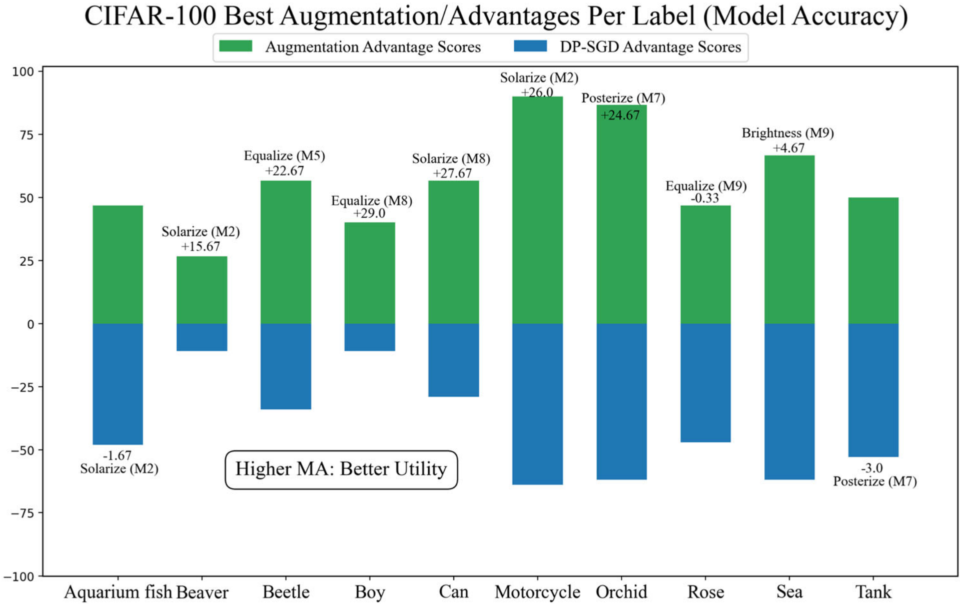


Fig. 15 CIFAR-100 attack accuracy advantages

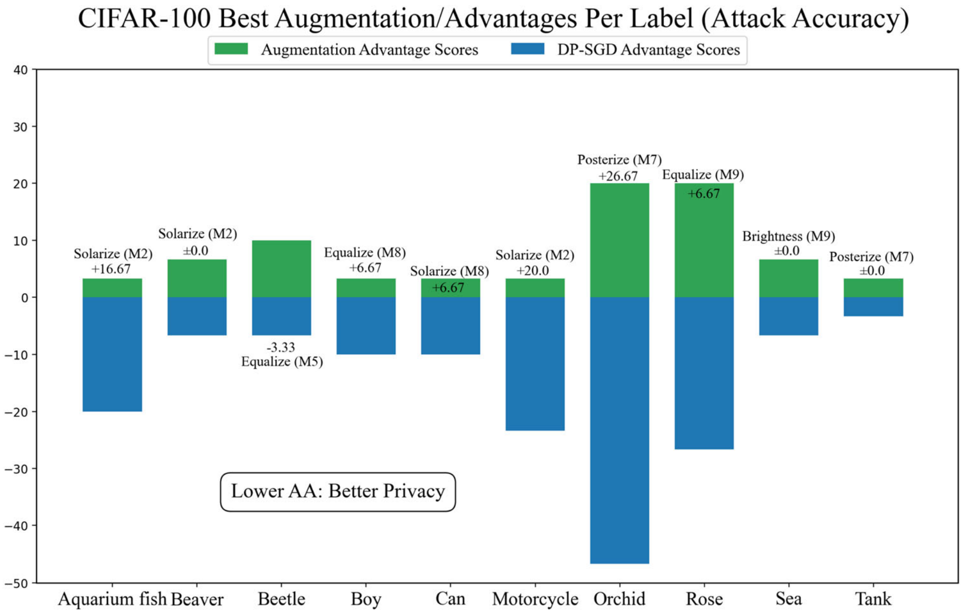
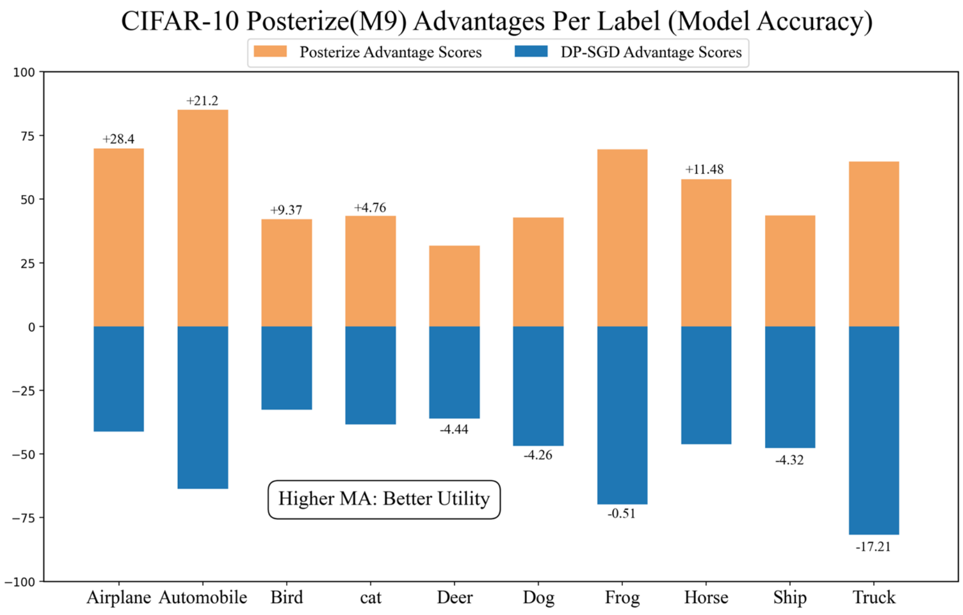


Table 6 Sampled advantage scores of CIFAR-100 augmentations

Labels Augmentation	Aquarium fish Solarize (M2)	Beaver Solarize (M2)	Beetle Equalize (M5)	Boy Equalize (M8)	Can Solarize (M8)
Advantage	15.34	15.67	19.34	35.67	34.34

Labels Augmentation	Motorcycle Solarize (M2)	Orchid Posterize (M7)	Rose Equalize (M9)	Sea Brightness (M9)	Tank Posterize (M7)
Advantage	46.0	51.34	6.34	4.67	-3.0

Fig. 16 CIFAR-10 posterize (M9) model accuracy advantages



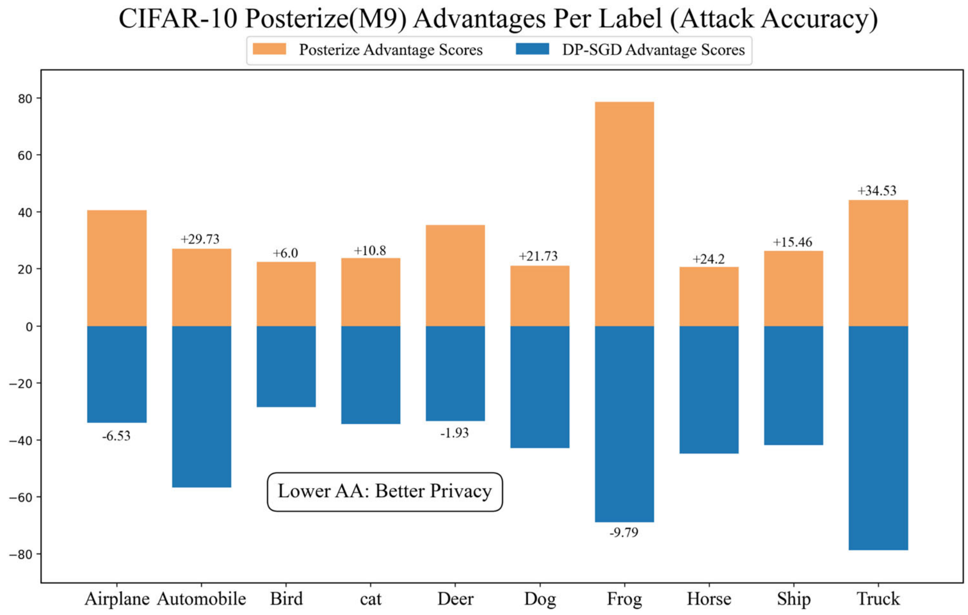


Fig. 17 CIFAR-10 posterize (M9) attack accuracy advantages

Table 7 CIFAR-10 advantage scores of posterize (M9)

Augmentation	Posterize (M9)									
Labels	Airplane	Automobile	Bird	Cat	Deer	Dog	Frog	Horse	Ship	Truck
Advantage	21.87	50.93	15.37	15.56	− 6.37	17.47	− 10.3	35.68	11.14	17.32

augmentation with magnitude 3 is applied. Figures 11 and 12 provide the model advantage and attack the advantage of CIFAR-10. The sum of values for each label provided in Figs. 11 and 12 is the total advantage scores provided in Table 5. A few outstanding results in Table 5 are automobile and truck labels that return very high advantage scores, 57.19 and 40.79 relatively. Both labels show the best performance when solarize augmentation with magnitude 7 is applied. The truck label shows lower model accuracy compared to DP-SGD results, however, it provides 59.33% lower attack accuracy, which is outstanding defense performance.

Same as the CIFAR-10, Figs. 13, 14, and 15 provide the advantage score of CIFAR-100 denoted in Table 6. We sampled the 10 best labels of CIFAR-100 due to the significant amount of classes in CIFAR-100. As mentioned before, color space augmentations return high advantage scores in most cases. In the case of the Orchid label, it returns a significant advantage score of 51.34, which is the result when posterize augmentation with magnitude 7 is applied. Note that CIFAR-100 differential privacy results return relatively low accuracy compared to CIFAR-10 due to the sensitivity towards the noise.

In this experiment, we found that posterize augmentation shows significant efficiency for the majority of introduced labels. All image data given in the dataset consists of red, green, and blue images, which are included in the RGB channel. Each channel consists of 8 bits and they support up to $2^8 = 256$ colors each. As the magnitude determines the number of bits to be removed from each color channel, only a few color pixels remain with high magnitude settings, therefore only the silhouette remains when reconstructed. The advantage scores of posterize augmentation are visualized in Figs. 16 and 17, and also the overall advantage score is provided in Table 7.

Finally, Fig. 18 shows the reconstructed images to visualize how to posterize augmentation prevents the data leakage from model inversion attack in the given environment. The overall results of CIFAR-10 are in the appendix section, and CIFAR-100 results are available from our website.[1]

6 Future works

We presented the results of our augmentation-based defense strategy against privacy attacks that reconstructs training data from model parameters. A few outstanding

[1] http://i2s.kennesaw.edu/resources.html.

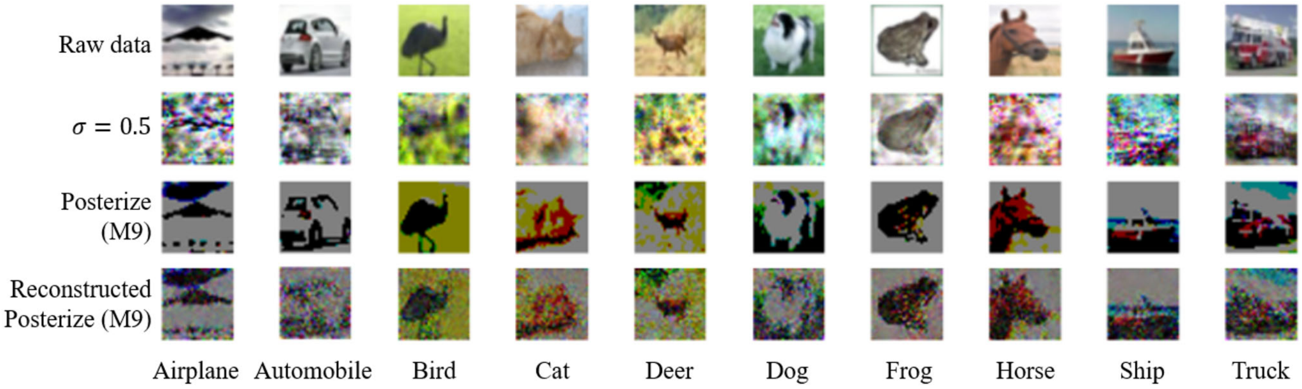


Fig. 18 Reconstructed images with Gaussian noise/Posterize (M9)

augmentations with optimized magnitude were found in the experiment, however, all the searching process was done manually this time, so there are many other outstanding augmentations that can give even better results compared to differential privacy-based defense strategies. These techniques will be analyzed and implemented in our future work. As we proved that distorting the given image through augmentation can prevent reconstruction from model parameters, future research will be about developing the adaptive augmentation that provides noticeable accuracy in classification while preventing the reconstruction of data. Our future research plan includes finding an automated solution that selects and applies the best augmentations to the given training data. For example, when an image that has 256 × 256 sizes are given as input, 65,536 pixels and a total 256^3 RGB value for each pixel will be observed. Based on the observation, selecting the best action from numerous cases, which means selecting and applying augmentation to the best pixels that affect the result of the reconstruction attack, will be the key to future research.

7 Conclusions

In this paper, we discussed federated learning and inherent privacy risks regarding the reconstruction of hidden training data from model parameters used in the training process. A traditional way to protect sensitive data from the proposed attacks that exploit the model parameter is to deploy differential privacy with a sufficient amount of noise. However, a few issues for deploying differential privacy exist, which are controlling the amount of noise and optimizing the hyperparameters. To present a new privacy-preserving solution that outperforms differential privacy with a simple implementation process, we conducted multiple reconstruction experiments applying 14 augmentations with 9 magnitudes. approximately 780,000 images were reconstructed during the experiment to secure the meaningful amount of data, and as a result, our experiment showed that a few augmentations successfully preserved the privacy against attacks exploiting model parameters and achieves noticeable accuracy in classification compared to differential privacy based defense strategy. We found a few good matches of augmentations and data classes from both datasets that returns the best performance during the experiment. Color space augmentations proposed in this paper shows superior performance to geometric transformations, and the posterize augmentation with the highest magnitude worked greatly for various image classes in both CIFAR datasets. Our augmentation-based defense strategy is easy to implement and can be applied regularly to whole data, therefore contributing to building a secure environment against model inversion attacks. Although the optimized augmentations and magnitudes for each label of the dataset were chosen manually this time, the adaptive augmentation algorithms and the optimized hyperparameters that outperform the current results will be found in the next research based on deep reinforcement learning.

Appendix

See Figs. 19, 20, 21, 22.

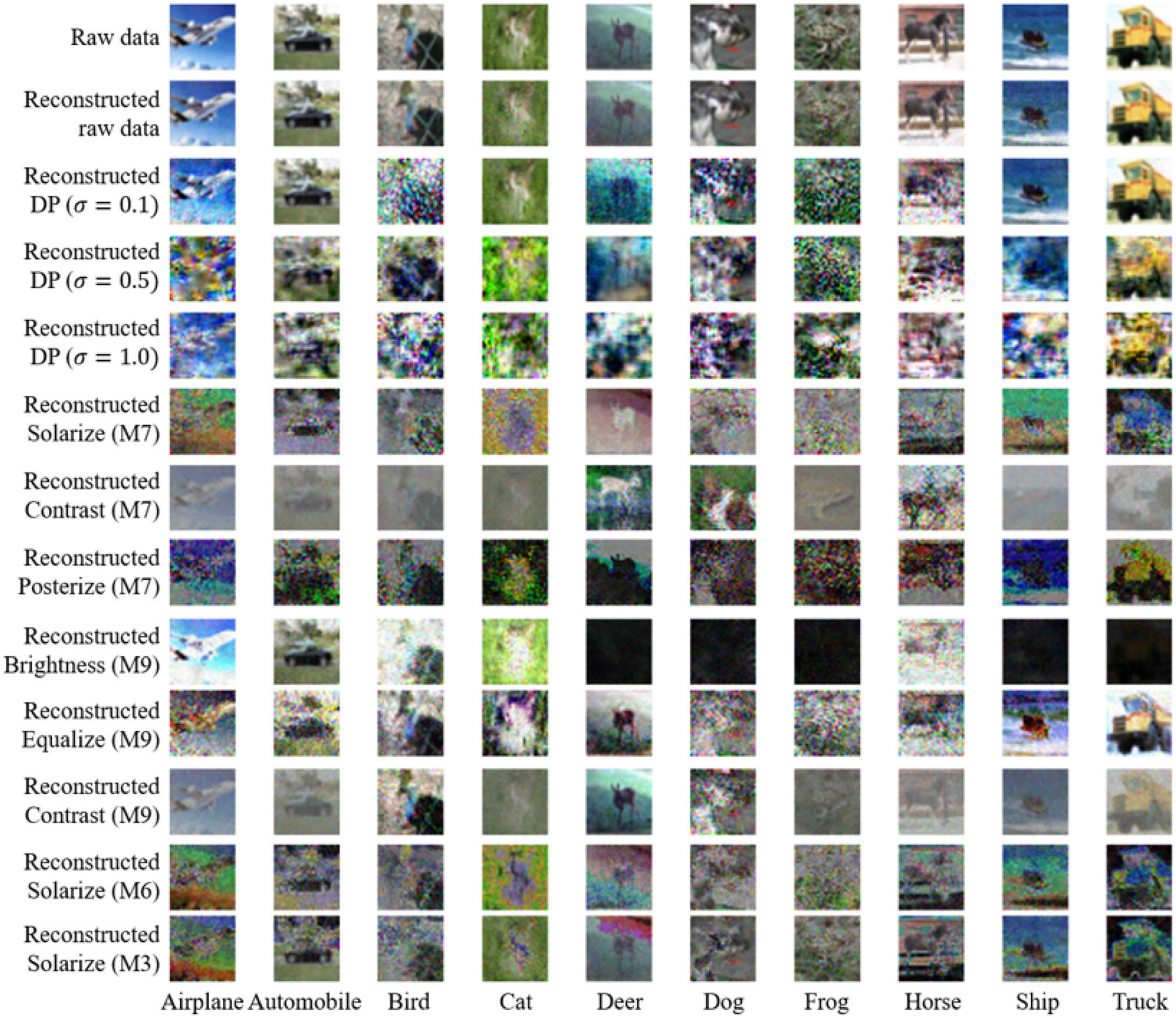
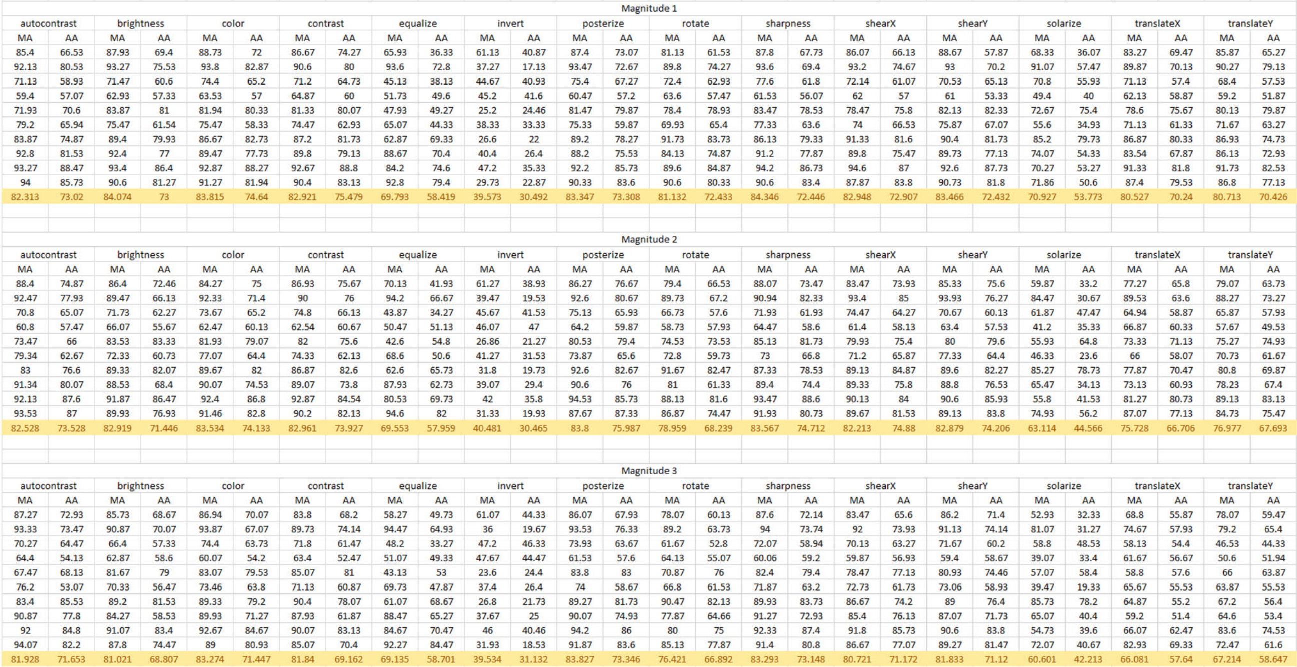

Raw data
Reconstructed raw data
Reconstructed DP ($\sigma = 0.1$)
Reconstructed DP ($\sigma = 0.5$)
Reconstructed DP ($\sigma = 1.0$)
Reconstructed Solarize (M7)
Reconstructed Contrast (M7)
Reconstructed Posterize (M7)
Reconstructed Brightness (M9)
Reconstructed Equalize (M9)
Reconstructed Contrast (M9)
Reconstructed Solarize (M6)
Reconstructed Solarize (M3)
Airplane
Automobile
Bird
Cat
Deer
Dog
Frog
Horse
Ship
Truck

Fig. 19 Notable visualization results of defense strategies



Magnitude 1

autocontrast		brightness		color		contrast		equalize		invert		posterize		rotate		sharpness		shearX		shearY		solarize		translateX		translateY	
MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA
85.4	66.53	87.93	69.4	88.73	72	86.67	74.27	65.93	36.33	61.13	40.87	87.4	73.07	81.13	61.53	87.8	67.73	86.07	66.13	88.67	57.87	68.33	36.07	83.27	69.47	85.87	65.27
92.13	80.53	93.27	75.53	93.8	82.87	90.6	80	93.6	72.8	37.27	17.13	93.47	72.67	89.8	74.27	93.6	69.4	93.2	74.67	93	70.2	91.07	57.47	89.87	70.13	90.27	79.13
71.13	58.93	71.47	60.6	74.4	65.2	71.2	64.73	45.13	38.13	44.67	40.93	75.4	67.27	72.4	62.93	77.6	61.8	72.14	61.07	70.53	65.13	70.8	55.93	71.13	57.4	68.4	57.53
59.4	57.07	62.93	57.33	63.53	57	64.87	60	51.73	49.6	45.2	41.6	60.47	57.2	63.6	57.47	61.53	56.07	62	57	61	53.33	49.4	40	62.13	58.87	59.2	51.87
71.93	70.6	83.87	81	81.94	80.33	81.33	80.07	47.93	49.27	25.2	24.46	81.47	79.87	78.4	78.93	83.47	78.53	78.47	75.8	82.13	82.33	72.67	75.4	78.6	75.67	80.13	79.87
79.2	65.94	75.47	61.54	75.47	58.33	74.47	62.93	65.07	44.33	38.33	33.33	75.33	59.87	69.93	65.4	77.33	63.6	74	66.53	75.87	67.07	55.6	34.93	71.13	61.33	71.67	63.27
83.87	74.87	89.4	79.93	86.67	82.73	87.2	81.73	62.87	69.33	26.6	22	89.2	78.27	91.73	83.73	86.13	79.33	91.33	81.6	90.4	81.73	85.2	79.73	86.87	80.33	86.93	74.73
92.8	81.53	92.4	77	89.47	77.73	89.8	79.13	88.67	70.4	40.4	26.4	88.2	75.53	84.13	74.87	91.2	77.87	89.8	75.47	89.73	77.13	74.07	54.33	83.54	67.87	86.13	72.93
93.27	88.47	93.4	86.4	92.87	88.27	92.67	88.8	84.2	74.6	47.2	35.33	92.2	85.73	89.6	84.87	94.2	86.73	94.6	87	92.6	87.73	70.27	53.27	91.33	81.8	91.73	82.53
94	85.73	90.6	81.27	91.27	81.94	90.4	83.13	92.8	79.4	29.73	22.87	90.33	83.6	90.6	80.33	90.6	83.4	87.87	83.8	90.73	81.8	71.86	50.6	87.4	79.53	86.8	77.13
82.313	73.02	84.074	73	83.815	74.64	82.921	75.479	69.793	58.419	39.573	30.492	83.347	73.308	81.132	72.433	84.346	72.446	82.948	72.907	83.466	72.432	70.927	53.773	80.527	70.24	80.713	70.426

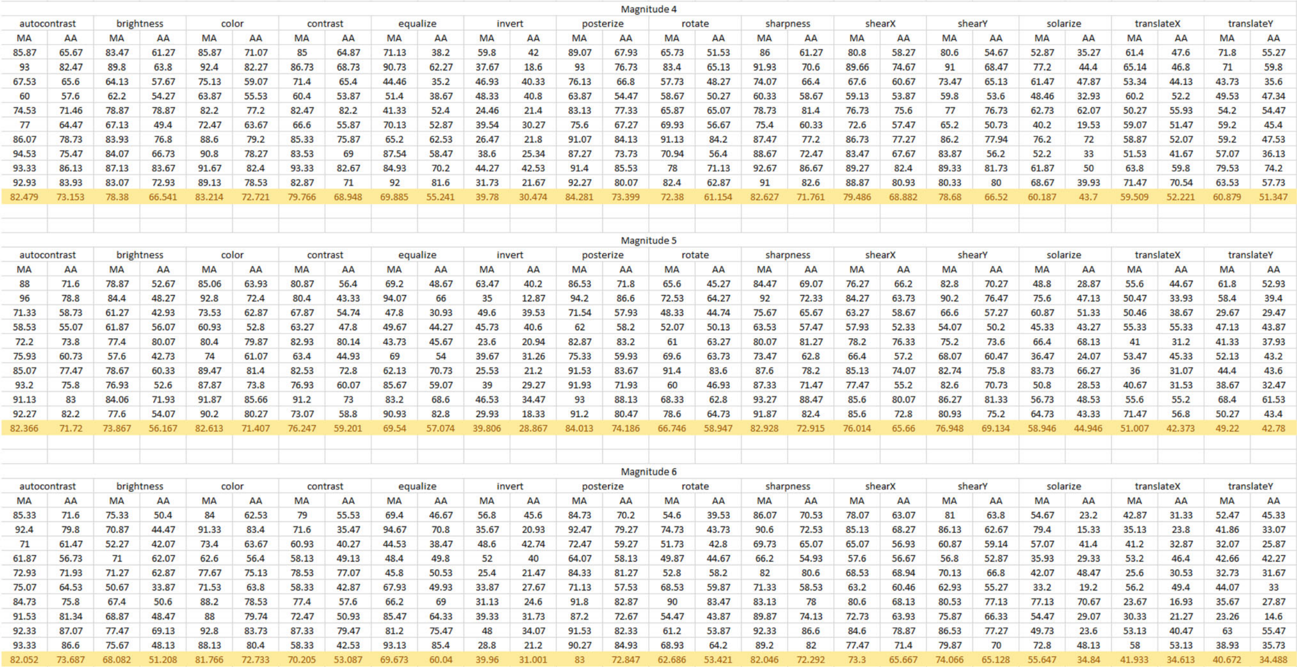
Magnitude 2

autocontrast		brightness		color		contrast		equalize		invert		posterize		rotate		sharpness		shearX		shearY		solarize		translateX		translateY	
MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA
88.4	74.87	86.4	72.46	84.27	75	86.93	75.67	70.13	41.93	61.27	38.93	86.27	76.67	79.4	66.53	88.07	73.47	83.47	73.93	85.33	75.6	59.87	33.2	77.27	65.8	79.07	63.73
92.47	77.93	89.47	66.13	92.33	71.4	90	76	94.2	66.67	39.47	19.53	92.6	80.67	89.73	67.2	90.94	82.33	93.4	85	93.93	76.27	84.47	30.67	89.53	63.6	88.27	73.27
70.8	65.07	71.73	62.27	73.67	65.2	74.8	66.13	43.87	34.27	45.67	41.53	75.13	65.93	66.73	57.6	71.93	61.93	74.47	64.27	70.67	60.13	61.87	47.47	64.94	58.87	65.87	57.93
60.8	57.47	66.07	55.67	62.47	60.13	62.54	60.67	50.47	51.13	46.07	47	64.2	59.87	58.73	57.93	64.47	58.6	61.4	58.13	63.4	57.53	41.2	35.33	66.87	60.33	57.67	49.53
73.47	66	83.53	83.33	81.93	79.07	82	75.6	42.6	54.8	26.86	21.27	80.53	79.4	74.53	73.53	85.13	81.73	79.93	75.4	80	79.6	55.93	64.8	73.33	71.13	75.27	74.93
79.34	62.67	72.33	60.73	77.07	64.4	74.33	62.13	68.6	50.6	41.27	31.53	73.87	65.6	72.8	59.73	73	66.8	71.2	65.87	77.33	64.4	46.33	23.6	66	58.07	70.73	61.67
83	76.6	89.33	82.07	89.67	82	86.87	82.6	62.6	65.73	31.8	19.73	92.6	82.67	91.67	82.47	87.33	78.53	89.13	84.87	89.6	82.27	85.27	78.73	77.87	70.47	80.8	69.87
91.34	80.07	88.53	68.4	90.07	74.53	89.07	73.8	87.93	62.73	39.07	29.4	90.6	76	81	61.33	89.4	74.4	89.33	75.8	88.8	76.53	65.47	34.13	73.13	60.93	78.23	67.4
92.13	87.6	91.87	86.47	92.4	86.8	92.87	84.54	80.53	69.73	42	35.8	94.53	85.73	88.13	81.6	93.47	88.6	90.13	84	90.6	85.93	55.8	41.53	81.27	80.73	89.13	83.13
93.53	87	89.93	76.93	91.46	82.8	90.2	82.13	94.6	82	31.33	19.93	87.67	87.33	86.87	74.47	91.93	80.73	89.67	81.53	89.13	83.8	74.93	56.2	87.07	77.13	84.73	75.47
82.528	73.528	82.919	71.446	83.534	74.133	82.961	73.927	69.553	57.959	40.481	30.465	83.8	75.987	78.959	68.239	83.567	74.712	82.213	74.88	82.879	74.206	63.114	44.566	75.728	66.706	76.977	67.693

Magnitude 3

autocontrast		brightness		color		contrast		equalize		invert		posterize		rotate		sharpness		shearX		shearY		solarize		translateX		translateY	
MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA
87.27	72.93	85.73	68.67	86.94	70.07	83.8	68.2	58.27	49.73	61.07	44.33	86.07	67.93	78.07	60.13	87.6	72.14	83.47	65.6	86.2	71.4	52.93	32.33	68.8	55.87	78.07	59.47
93.33	73.47	90.87	70.07	93.87	67.07	89.73	74.14	94.47	64.93	36	19.67	93.53	76.33	89.2	63.73	94	73.74	92	73.93	91.13	74.14	81.07	31.27	74.67	57.93	79.2	65.4
70.27	64.47	66.4	57.33	74.4	63.73	71.8	61.47	48.2	33.27	47.2	46.33	73.93	63.67	61.67	52.8	72.07	58.94	70.13	63.27	71.67	60.2	58.8	48.53	58.13	54.4	46.53	44.33
64.4	54.13	62.87	58.6	60.07	54.2	63.4	52.47	51.07	49.33	47.67	44.47	61.53	57.6	64.13	55.07	60.06	59.2	59.87	56.93	59.4	58.67	39.07	33.4	61.67	56.67	50.6	51.94
67.47	68.13	81.67	79	83.07	79.53	85.07	81	43.13	53	23.6	24.4	83.8	83	70.87	76	82.4	79.4	78.47	77.13	80.93	74.46	57.07	58.4	58.8	57.6	66	63.87
76.2	53.07	70.33	56.47	73.46	63.8	71.13	60.87	69.73	47.87	37.4	26.4	74	58.67	66.8	61.53	71.87	63.2	72.73	61.73	73.06	58.93	39.47	19.33	65.67	55.53	63.87	55.53
83.4	85.53	89.2	81.53	89.33	79.2	90.4	78.07	61.07	68.67	26.8	21.73	89.27	81.73	90.47	82.13	89.93	83.73	86.67	74.2	89	76.4	85.73	78.2	64.87	55.2	67.2	56.4
90.87	77.8	84.27	58.53	89.93	71.27	87.93	61.87	88.47	65.27	37.67	25	90.07	74.93	77.87	64.66	91.27	72.93	85.4	76.13	87.07	71.73	65.07	40.4	59.2	51.4	64.6	53.4
92	84.8	91.07	83.4	92.67	84.67	90.07	83.13	84.67	70.47	46	40.46	94.2	86	80	75	92.33	87.4	91.8	85.73	90.6	83.8	54.73	39.6	66.07	62.47	83.6	74.53
94.07	82.2	87.8	74.47	89	80.93	85.07	70.4	92.27	84.47	31.93	18.53	91.87	83.6	85.13	77.87	91.4	80.8	86.67	77.07	89.27	81.47	72.07	40.67	82.93	69.33	72.47	61.6
81.928	71.653	81.021	68.807	83.274	71.447	81.84	69.162	69.135	58.701	39.534	31.132	83.827	73.346	76.421	66.892	83.293	73.148	80.721	71.172	81.833	71.12	60.601	42.213	66.081	57.64	67.214	58.647

Fig. 20 Accuracy table for augmentations with magnitude 1 to 3



Magnitude 4

autocontrast		brightness		color		contrast		equalize		invert		posterize		rotate		sharpness		shearX		shearY		solarize		translateX		translateY	
MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA
85.87	65.67	83.47	61.27	85.87	71.07	85	64.87	71.13	38.2	59.8	42	89.07	67.93	65.73	51.53	86	61.27	80.8	58.27	80.6	54.67	52.87	35.27	61.4	47.6	71.8	55.27
93	82.47	89.8	63.8	92.4	82.27	86.73	68.73	90.73	62.27	37.67	18.6	93	76.73	83.4	65.13	91.93	70.6	89.66	74.67	91	68.47	77.2	44.4	65.14	46.8	71	59.8
67.53	65.6	64.13	57.67	75.13	59.07	71.4	65.4	44.46	35.2	46.93	40.33	76.13	66.8	57.73	48.27	74.07	66.4	67.6	60.67	73.47	65.13	61.47	47.87	53.34	44.13	43.73	35.6
60	57.6	62.2	54.27	63.87	55.53	60.4	53.87	51.4	38.67	48.33	40.8	63.87	54.47	58.67	50.27	60.33	58.67	59.13	53.87	59.8	53.6	48.46	32.93	60.2	52.2	49.53	47.34
74.53	71.46	78.87	78.87	82.2	77.2	82.47	82.2	41.33	52.4	24.46	21.4	83.13	77.33	65.87	65.07	78.73	81.4	76.73	75.6	77	76.73	62.73	62.07	50.27	55.93	54.2	54.47
77	64.47	67.13	49.4	72.47	63.67	66.6	55.87	70.13	52.87	39.54	30.27	75.6	67.27	69.93	56.67	75.4	60.33	72.6	57.47	65.2	50.73	40.2	19.53	59.07	51.47	59.2	45.4
86.07	78.73	83.93	76.8	88.6	79.2	85.33	75.87	65.2	62.53	26.47	21.8	91.07	84.13	91.13	84.2	87.47	77.2	86.73	77.27	86.2	77.94	76.2	72	58.87	52.07	59.2	47.53
94.53	75.47	84.07	66.73	90.8	78.27	83.53	69	87.54	58.47	38.6	25.34	87.27	73.73	70.94	56.4	88.67	72.47	83.47	67.67	83.87	56.2	52.2	33	51.53	41.67	57.07	36.13
93.33	86.13	87.13	83.67	91.67	82.4	93.33	82.67	84.93	70.2	44.27	42.53	91.4	85.53	78	71.13	92.67	86.67	89.27	82.4	89.33	81.73	61.87	50	63.8	59.8	79.53	74.2
92.93	83.93	83.07	72.93	89.13	78.53	82.87	71	92	81.6	31.73	21.67	92.27	80.07	82.4	62.87	91	82.6	88.87	80.93	80.33	80	68.67	39.93	71.47	70.54	63.53	57.73
82.479	73.153	78.38	66.541	83.214	72.721	79.766	68.948	69.885	55.241	39.78	30.474	84.281	73.399	72.38	61.154	82.627	71.761	79.486	68.882	78.68	66.52	60.187	43.7	59.509	52.221	60.879	51.347

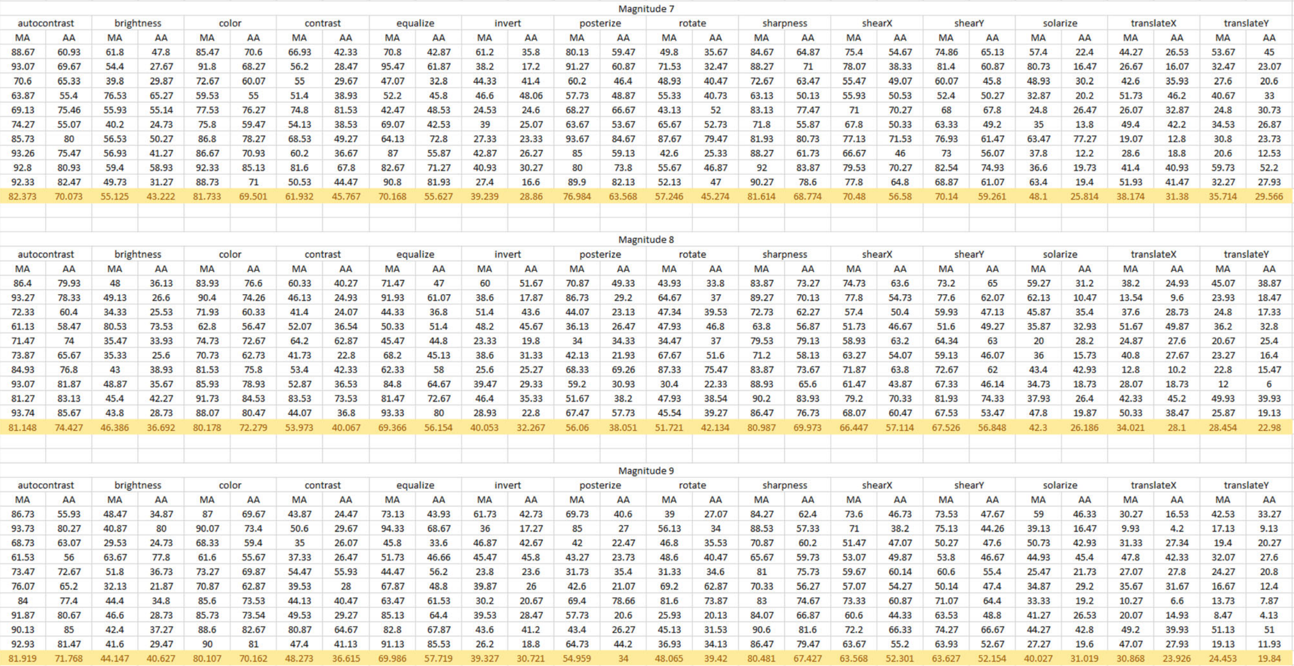
Magnitude 5

autocontrast		brightness		color		contrast		equalize		invert		posterize		rotate		sharpness		shearX		shearY		solarize		translateX		translateY	
MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA
88	71.6	78.87	52.67	85.06	63.93	80.87	56.4	69.2	48.67	63.47	40.2	86.53	71.8	65.6	45.27	84.47	69.07	76.27	66.2	82.8	70.27	48.8	28.87	55.6	44.67	61.8	52.93
96	78.8	84.4	48.27	92.8	72.4	80.4	43.33	94.07	66	35	12.87	94.2	86.6	72.53	64.27	92	72.33	84.27	63.73	90.2	76.47	75.6	47.13	50.47	33.93	58.4	39.4
71.33	58.73	61.27	42.93	73.53	62.87	67.87	54.74	47.8	30.93	49.6	39.53	71.54	57.93	48.33	44.74	75.67	65.67	63.27	58.67	66.6	57.27	60.87	51.33	50.46	38.67	29.67	29.47
58.53	55.07	61.87	56.07	60.93	52.8	63.27	47.8	49.67	44.27	45.73	40.6	62	58.2	52.07	50.13	63.53	57.47	57.93	52.33	54.07	50.2	45.33	43.27	55.33	55.33	47.13	43.87
72.2	73.8	77.4	80.07	80.4	79.87	82.93	80.14	43.73	45.67	23.6	20.94	82.87	83.2	61	63.27	80.07	81.27	78.2	76.33	75.2	73.6	66.4	68.13	41	31.2	41.33	37.93
75.93	60.73	57.6	42.73	74	61.07	63.4	44.93	69	54	39.67	31.26	75.33	59.93	69.6	63.73	73.47	62.8	66.4	57.2	68.07	60.47	36.47	24.07	53.47	45.33	52.13	43.2
85.07	77.47	78.67	60.33	89.47	81.4	82.53	72.8	62.13	70.73	25.53	21.2	91.53	83.67	91.4	83.6	87.6	78.2	85.13	74.07	82.74	75.8	83.73	66.27	36	31.07	44.4	43.6
93.2	75.8	76.93	52.6	87.87	73.8	76.93	60.07	85.67	59.07	39	29.27	91.93	71.93	60	46.93	87.33	71.47	77.47	55.2	82.6	70.73	50.8	28.53	40.67	31.53	38.67	32.47
91.13	83	84.06	71.93	91.87	85.66	91.2	73	83.2	68.6	46.53	34.47	93	88.13	68.33	62.8	93.27	88.47	85.6	80.07	86.27	81.33	56.73	48.53	55.6	55.2	68.4	61.53
92.27	82.2	77.6	54.07	90.2	80.27	73.07	58.8	90.93	82.8	29.93	18.33	91.2	80.47	78.6	64.73	91.87	82.4	85.6	72.8	80.93	75.2	64.73	43.33	71.47	56.8	50.27	43.4
82.366	71.72	73.867	56.167	82.613	71.407	76.247	59.201	69.54	57.074	39.806	28.867	84.013	74.186	66.746	58.947	82.928	72.915	76.014	65.66	76.948	69.134	58.946	44.946	51.007	42.373	49.22	42.78

Magnitude 6

autocontrast		brightness		color		contrast		equalize		invert		posterize		rotate		sharpness		shearX		shearY		solarize		translateX		translateY	
MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA
85.33	71.6	75.33	50.4	84	62.53	79	55.53	69.4	46.67	56.8	45.6	84.73	70.2	54.6	39.53	86.07	70.53	78.07	63.07	81	63.8	54.67	23.2	42.87	31.33	52.47	45.33
92.4	79.8	70.87	44.47	91.33	83.4	71.6	35.47	94.67	70.8	35.67	20.93	92.47	79.27	74.73	43.73	90.6	72.53	85.13	68.27	86.13	62.67	79.4	15.33	35.13	23.8	41.86	33.07
71	61.47	52.27	42.07	73.4	63.67	60.93	40.27	44.53	38.47	48.6	42.74	72.47	59.27	51.73	42.8	69.73	65.07	65.07	56.93	60.87	59.14	57.07	41.4	41.2	32.87	32.07	25.87
61.87	56.73	71	62.07	62.6	56.4	58.13	49.13	48.4	49.8	52	40	64.07	58.13	49.87	44.67	66.2	54.93	57.6	56.67	56.8	52.87	35.93	29.33	53.2	46.4	42.66	42.27
72.93	71.93	71.27	62.87	77.67	75.13	78.53	77.07	45.8	50.53	25.4	21.47	84.33	81.27	52.8	58.2	82	80.6	68.53	68.94	70.13	66.8	42.07	48.47	25.6	30.53	32.73	31.67
75.07	64.53	50.67	33.87	71.53	63.8	58.33	42.87	67.93	49.93	33.87	27.67	71.13	57.53	68.53	59.87	71.33	58.53	63.2	60.46	62.93	55.27	33.2	19.2	56.2	49.4	44.07	33
84.73	75.8	67.4	50.6	88.2	78.53	77.4	57.6	66.2	69	31.13	24.6	91.8	82.87	90	83.47	83.13	78	80.6	68.13	80.53	77.13	77.13	70.67	23.67	16.93	35.67	27.87
91.53	81.34	68.87	48.47	88	79.74	72.47	50.93	85.47	64.33	39.33	31.73	87.2	72.67	54.47	43.87	89.87	74.13	72.73	63.93	75.87	66.33	54.47	29.07	30.33	21.27	23.26	14.6
92.33	87.07	77.47	69.13	92.8	83.73	87.33	79.47	81.2	75.47	48	34.07	91.53	82.33	61.2	53.87	92.33	86.6	84.6	78.87	86.53	77.27	49.73	23.6	53.13	40.47	63	55.47
93.33	86.6	75.67	48.13	88.13	80.4	58.33	42.53	93.13	85.4	28.8	21.2	90.27	84.93	68.93	64.2	89.2	82	77.47	71.4	79.87	70	72.8	48.13	58	53.13	38.93	35.73
82.052	73.687	68.082	51.208	81.766	72.733	70.205	53.087	69.673	60.04	39.96	31.001	83	72.847	62.686	53.421	82.046	72.292	73.3	65.667	74.066	65.128	55.647	34.84	41.933	34.613	40.672	34.488

Fig. 21 Accuracy table for augmentations with magnitude 4 to 6



Magnitude 7

autocontrast		brightness		color		contrast		equalize		invert		posterize		rotate		sharpness		shearX		shearY		solarize		translateX		translateY	
MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA
88.67	60.93	61.8	47.8	85.47	70.6	66.93	42.33	70.8	42.87	61.2	35.8	80.13	59.47	49.8	35.67	84.67	64.87	75.4	54.67	74.86	65.13	57.4	22.4	44.27	26.53	53.67	45
93.07	69.67	54.4	27.67	91.8	68.27	56.2	28.47	95.47	61.87	38.2	17.2	91.27	60.87	71.53	32.47	88.27	71	78.07	38.33	81.4	60.87	80.73	16.47	26.67	16.07	32.47	23.07
70.6	65.33	39.8	29.87	72.67	60.07	55	29.67	47.07	32.8	44.33	41.4	60.2	46.4	48.93	40.47	72.67	63.47	55.47	49.07	60.07	45.8	48.93	30.2	42.6	35.93	27.6	20.6
63.87	55.4	76.53	65.27	59.53	55	51.4	38.93	52.2	45.8	46.6	48.06	57.73	48.87	55.33	40.73	63.13	50.13	55.93	50.53	52.4	50.27	32.87	20.2	51.73	46.2	40.67	33
69.13	75.46	55.93	55.14	77.53	76.27	74.8	81.53	42.47	48.53	24.53	24.6	68.27	66.67	43.13	52	83.13	77.47	71	70.27	68	67.8	24.8	26.47	26.07	32.87	24.8	30.73
74.27	55.07	40.2	24.73	75.8	59.47	54.13	38.53	69.07	42.53	39	25.07	63.67	53.67	65.67	52.73	71.8	55.87	67.8	50.33	63.33	49.2	35	13.8	49.4	42.2	34.53	26.87
85.73	80	56.53	50.27	86.8	78.27	68.53	49.27	64.13	72.8	27.33	23.33	93.67	84.67	87.67	79.47	81.93	80.73	77.13	71.53	76.93	61.47	63.47	77.27	19.07	12.8	30.8	23.73
93.26	75.47	56.93	41.27	86.67	70.93	60.2	36.67	87	55.87	42.87	26.27	85	59.13	42.6	25.33	88.27	61.73	66.67	46	73	56.07	37.8	12.2	28.6	18.8	20.6	12.53
92.8	80.93	59.4	58.93	92.33	85.13	81.6	67.8	82.67	71.27	40.93	30.27	80	73.8	55.67	46.87	92	83.87	79.53	70.27	82.54	74.93	36.6	19.73	41.4	40.93	59.73	52.2
92.33	82.47	49.73	31.27	88.73	71	50.53	44.47	90.8	81.93	27.4	16.6	89.9	82.13	52.13	47	90.27	78.6	77.8	64.8	68.87	61.07	63.4	19.4	51.93	41.47	32.27	27.93
82.373	70.073	55.125	43.222	81.733	69.501	61.932	45.767	70.168	55.627	39.239	28.86	76.984	63.568	57.246	45.274	81.614	68.774	70.48	56.58	70.14	59.261	48.1	25.814	38.174	31.38	35.714	29.566

Magnitude 8

autocontrast		brightness		color		contrast		equalize		invert		posterize		rotate		sharpness		shearX		shearY		solarize		translateX		translateY	
MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA
86.4	79.93	48	36.13	83.93	76.6	60.33	40.27	71.47	47	60	51.67	70.87	49.33	43.93	33.8	83.87	73.27	74.73	63.6	73.2	65	59.27	31.2	38.2	24.93	45.07	38.87
93.27	78.33	49.13	26.6	90.4	74.26	46.13	24.93	91.93	61.07	38.6	17.87	86.73	29.2	64.67	37	89.27	70.13	77.8	54.73	77.6	62.07	62.13	10.47	13.54	9.6	23.93	18.47
72.33	60.4	34.33	25.53	71.93	60.33	41.4	24.07	44.33	36.8	51.4	43.6	44.07	23.13	47.34	39.53	72.73	62.27	57.4	50.4	59.93	47.13	45.87	35.4	37.6	28.73	24.8	17.33
61.13	58.47	80.53	73.53	62.8	56.47	52.07	36.54	50.33	51.4	48.2	45.67	36.13	26.47	47.93	46.8	63.8	56.87	51.73	46.67	51.6	49.27	35.87	32.93	51.67	49.87	36.2	32.8
71.47	74	35.47	33.93	74.73	72.67	64.2	62.87	45.47	44.8	23.33	19.8	34	34.33	34.47	37	79.53	79.13	58.93	63.2	64.34	63	20	28.2	24.87	27.6	20.67	25.4
73.87	65.67	35.33	25.6	70.73	62.73	41.73	22.8	68.2	45.13	38.6	31.33	42.13	21.93	67.67	51.6	71.2	58.13	63.27	54.07	59.13	46.07	36	15.73	40.8	27.67	23.27	16.4
84.93	76.8	43	38.93	81.53	75.8	53.4	42.33	62.33	58	25.6	25.27	68.33	69.26	87.33	75.47	83.87	73.67	71.87	63.8	72.67	62	43.4	42.93	12.8	10.2	22.8	15.47
93.07	81.87	48.87	35.67	85.93	78.93	52.87	36.53	84.8	64.67	39.47	29.33	59.2	30.93	30.4	22.33	88.93	65.6	61.47	43.87	67.33	46.14	34.73	18.73	28.07	18.73	12	6
81.27	83.13	45.4	42.27	91.73	84.53	83.53	73.53	81.47	72.67	46.4	35.33	51.67	38.2	47.93	38.54	90.2	83.93	79.2	70.33	81.93	74.33	37.93	26.4	42.33	45.2	49.93	39.93
93.74	85.67	43.8	28.73	88.07	80.47	44.07	36.8	93.33	80	28.93	22.8	67.47	57.73	45.54	39.27	86.47	76.73	68.07	60.47	67.53	53.47	47.8	19.87	50.33	38.47	25.87	19.13
81.148	74.427	46.386	36.692	80.178	72.279	53.973	40.067	69.366	56.154	40.053	32.267	56.06	38.051	51.721	42.134	80.987	69.973	66.447	57.114	67.526	56.848	42.3	26.186	34.021	28.1	28.454	22.98

Magnitude 9

autocontrast		brightness		color		contrast		equalize		invert		posterize		rotate		sharpness		shearX		shearY		solarize		translateX		translateY	
MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA	MA	AA
86.73	55.93	48.47	34.87	87	69.67	43.87	24.47	73.13	43.93	61.73	42.73	69.73	40.6	39	27.07	84.27	62.4	73.6	46.73	73.53	47.67	59	46.33	30.27	16.53	42.53	33.27
93.73	80.27	40.87	80	90.07	73.4	50.6	29.67	94.33	68.67	36	17.27	85	27	56.13	34	88.53	57.33	71	38.2	75.13	44.26	39.13	16.47	9.93	4.2	17.13	9.13
68.73	63.07	29.53	24.73	68.33	59.4	35	26.07	45.8	33.6	46.87	42.67	42	22.47	46.8	35.53	70.87	60.2	51.47	47.07	50.27	47.6	50.73	42.93	31.33	27.34	19.4	20.27
61.53	56	63.67	77.8	61.6	55.67	37.33	26.47	51.73	46.66	45.47	45.8	43.27	23.73	48.6	40.47	65.67	59.73	53.07	49.87	53.8	46.67	44.93	45.4	47.8	42.33	32.07	27.6
73.47	72.67	51.8	36.73	73.27	69.87	54.47	55.93	44.47	56.2	23.8	23.6	31.73	35.4	31.33	34.6	81	75.73	59.67	60.14	60.6	55.4	25.47	21.73	27.07	27.8	24.27	20.8
76.07	65.2	32.13	21.87	70.87	62.87	39.53	28	67.87	48.8	39.87	26	42.6	21.07	69.2	62.87	70.33	56.27	57.07	54.27	50.14	47.4	34.87	29.2	35.67	31.67	16.67	12.4
84	77.4	44.4	34.8	85.6	73.53	44.13	40.47	63.47	61.53	30.2	20.67	69.4	78.66	81.6	73.87	83	74.67	73.33	60.87	71.07	64.4	33.33	19.2	10.27	6.6	13.73	7.87
91.87	80.67	46.6	28.73	85.73	73.54	49.53	29.27	85.13	64.4	39.53	28.47	57.73	20.6	25.93	20.13	84.07	66.87	60.6	44.33	63.53	48.8	41.27	26.53	20.07	14.93	8.47	4.13
90.13	85	42.4	37.27	88.6	82.67	80.87	64.67	82.8	67.87	43.6	41.2	43.4	26.27	45.13	31.53	90.6	81.6	72.2	66.33	74.27	66.67	44.27	42.8	49.2	39.93	51.13	51
92.93	81.47	41.6	29.47	90	81	47.4	41.13	91.13	85.53	26.2	18.8	64.73	44.2	36.93	34.13	86.47	79.47	63.67	55.2	63.93	52.67	27.27	19.6	47.07	27.93	19.13	11.93
81.919	71.768	44.147	40.627	80.107	70.162	48.273	36.615	69.986	57.719	39.327	30.721	54.959	34	48.065	39.42	80.481	67.427	63.568	52.301	63.627	52.154	40.027	31.019	30.868	23.926	24.453	19.84

Fig. 22 Accuracy table for augmentations with magnitude 7 to 9

Author contributions JS contributed to the study's conception and design. Funding acquisitions were performed by JS and KK. Material preparation, data collection, and analysis were performed by SS. The first draft of the manuscript was written by SS. MB and KS proofread and commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding This research was supported by the MSIT (Ministry of Science, ICT), Korea, under the High-Potential Individuals Global Training Program (2021-0-01547-001) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation).

Data availability The data that support the findings of this study are available from https://www.cs.toronto.edu/∼kriz/cifar.html.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethical approval This paper has never been submitted or introduced in any form to any journals or conferences before.

References

1. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. Artif. Intell. Statist. pp. 1273–1282 (2017)
2. Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership inference attacks against machine learning models. In: 2017 IEEE Symposium on Security and Privacy (SP), pp. 3–18. IEEE (2017)
3. Fredrikson, M., Lantz, E., Jha, S., Lin, S., Page, D., Ristenpart, T.: Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing. In: 23rd {USENIX} Security Symposium ({USENIX} Security 14), pp. 17–32 (2014)
4. Fredrikson, M., Jha, S., Ristenpart, T.: Model inversion attacks that exploit confidence information and basic countermeasures. In: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, pp. 1322–1333 (2015)
5. Hidano, S., Murakami, T., Katsumata, S., Kiyomoto, S., Hanaoka, G.: Model inversion attacks for prediction systems: without knowledge of non-sensitive attributes. In: 2017 15th Annual Conference on Privacy, Security and Trust (PST), pp. 115–11509. IEEE (2017)
6. Zhu, L., Han, S.: Deep leakage from gradients. In: Federated Learning, pp. 17–31. Springer (2020)
7. Zhao, B., Mopuri, K.R., Bilen, H.: idlg: Improved deep leakage from gradients. arXiv preprint arXiv:2001.02610 (2020)
8. Geiping, J., Bauermeister, H., Dröge, H., Moeller, M.: Inverting gradients–how easy is it to break privacy in federated learning? arXiv preprint arXiv:2003.14053 (2020)
9. Zhang, Y., Jia, R., Pei, H., Wang, W., Li, B., Song, D.: The secret revealer: generative model-inversion attacks against deep neural

networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 253–261 (2020)
10. He, Z., Zhang, T., Lee, R.B.: Model inversion attacks against collaborative inference. In: Proceedings of the 35th Annual Computer Security Applications Conference, pp. 148–162 (2019)
11. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating noise to sensitivity in private data analysis. In: Theory of Cryptography Conference, pp. 265–284. Springer (2006)
12. McSherry, F., Talwar, K.: Mechanism design via differential privacy. In: 48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07), pp. 94–103. IEEE (2007)
13. Shokri, R., Shmatikov, V.: Privacy-preserving deep learning. In: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, pp. 1310–1321 (2015)
14. Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L.: Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, pp. 308–318 (2016)
15. Phan, N., Wu, X., Hu, H., Dou, D.: Adaptive laplace mechanism: Differential privacy preservation in deep learning. In: 2017 IEEE International Conference on Data Mining (ICDM), pp. 385–394. IEEE (2017)
16. Mironov, I.: Rényi differential privacy. In: 2017 IEEE 30th Computer Security Foundations Symposium (CSF), pp. 263–275 (2017). IEEE
17. Truex, S., Liu, L., Chow, K.-H., Gursoy, M.E., Wei, W.: Ldp-fed: Federated learning with local differential privacy. In: Proceedings of the Third ACM International Workshop on Edge Systems, Analytics and Networking, pp. 61–66 (2020)
18. Girgis, A., Data, D., Diggavi, S., Kairouz, P., Suresh, A.T.: Shuffled model of differential privacy in federated learning. In: International Conference on Artificial Intelligence and Statistics, pp. 2521–2529. PMLR (2021)
19. DeVries, T., Taylor, G.W.: Improved regularization of convolutional neural networks with cutout. arXiv preprint arXiv:1708.04552 (2017)
20. Zhong, Z., Zheng, L., Kang, G., Li, S., Yang, Y.: Random erasing data augmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 13001–13008 (2020)
21. Wu, R., Yan, S., Shan, Y., Dang, Q., Sun, G.: Deep image: Scaling up image recognition. **7**(8) (2015). arXiv preprint arXiv:1501.02876
22. Zheng, Z., Zheng, L., Yang, Y.: Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3754–3762 (2017)
23. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. (2020) arXiv preprint arXiv:2011.13456
24. Lin, J., Li, Y., Yang, G.: Fpgan: face de-identification method with generative adversarial networks for social robots. Neural Netw. **133**, 132–147 (2021)
25. Wu, H., Zheng, S., Zhang, J., Huang, K.: Gp-gan: Towards realistic high-resolution image blending. In: Proceedings of the 27th ACM International Conference on Multimedia, pp. 2487–2495 (2019)
26. Choi, Y., Uh, Y., Yoo, J., Ha, J.-W.: Stargan v2: Diverse image synthesis for multiple domains. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8188–8197 (2020)
27. Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V., Le, Q.V.: Autoaugment: earning augmentation policies from data. (2018) arXiv preprint arXiv:1805.09501
28. Lim, S., Kim, I., Kim, T., Kim, C., Kim, S.: Fast autoaugment. arXiv preprint arXiv:1905.00397 (2019)
29. Hataya, R., Zdenek, J., Yoshizoe, K., Nakayama, H.: Faster autoaugment: learning augmentation strategies using backpropagation. In: European Conference on Computer Vision, pp. 1–16. Springer (2020)
30. Cubuk, E.D., Zoph, B., Shlens, J., Le, Q.V.: Randaugment: practical automated data augmentation with a reduced search space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp. 702–703 (2020)
31. Dwork, C., Roth, A., et al.: The algorithmic foundations of differential privacy. Found. Trends Theor. Compt. Sci. **9**(3–4), 211–407 (2014)
32. Shorten, C., Khoshgoftaar, T.M.: A survey on image data augmentation for deep learning. J. Big Data **6**(1), 1–48 (2019)
33. Waites, C.: Pyvacy: towards practical differential privacy for deep learning (2019)
34. Opacus PyTorch library. https://opacus.ai

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Seunghyeon Shin is a student at Kennesaw State University pursuing master's degree in Computer Science and a member of Information and Intelligent Security Lab. He received the Bachelor of Science degree in Computer Science from Kennesaw State University in 2019. His research interests include Deep Learning, Image processing, and Cybersecurity.

Mallika Boyapati received the B.Tech degree in Electronics and Computer Engineering from Koneru Lakshmaiah (KL) University, Vijayawada, India in 2016, and the M.Sc. degree in Applied Computer Science from Columbus State University, Columbus, Georgia, in 2018. From 2018 to 2021, she worked as a Senior Data Analyst at T-Mobile. She joined as a Ph.D. student in Analytics and Data Science at Kennesaw State University and is working as a research assistant in information and Intelligent Security (IIS) Lab. Her research interests include Data Analytics, Data Science, Machine Learning, and Cybersecurity.

Kun Suo received the BS degree in software engineering from the Nanjing University, China, in 2012, and Ph.D. degree from the University of Texas at Arlington in 2019. He is currently an Assistant Professor in the Department of Computer Science at Kennesaw State University. His research interests include the areas of cloud computing, virtualization, operating systems, Java virtual machines, software defined network, and machine learning systems. He is a member of the IEEE and ACM.

Kyungtae Kang received a B.S. degree in computer science and engineering, followed by M.S. and Ph.D. degrees in electrical engineering and computer science, from Seoul National University, Seoul, Korea, in 1999, 2001, and 2007, respectively. From 2008 to 2010, he was a postdoctoral research associate at the University of Illinois at Urbana-Champaign, IL, USA. In 2011, he joined the Department of Computer Science and Engineering at Hanyang University, where he is currently a tenured professor. His research interests lie primarily in systems, including operating systems, mobile systems, distributed systems, and real-time embedded systems. His recent research interest is in the interdisciplinary area of cyber-physical systems.

Junggab Son received the BSE degree in computer science and engineering from Hanyang University, Ansan, South Korea (2009), and the Ph.D. degree in computer science and engineering from Hanyang University, Seoul, South Korea (2014). From 2014 to 2016, he was a Post-doctoral Research Associate with the Department of Math and Physics, North Carolina Central University. From 2016 to 2018, he was a Research Fellow and a Limited-term Assistant Professor at Kennesaw State University. Since 2018, he has been an Assistant Professor of Computer Science and a Director of Information and Intelligent Security (IIS) Laboratory at Kennesaw State University. His research interests include applied cryptography, privacy preservation, blockchain and smart contract, malware detection, and security/privacy issues in artificial intelligent algorithms. He is a senior member of IEEE and a member of ACM.