A Data Augmentation-based Defense Method Against Adversarial Attacks in Neural Networks

Yi Zeng[1], Han Qiu[2], Gerard Memmi[2], and Meikang Qiu[3]

[1] University of California San Diego, CA, USA, 92122.
y4zeng@eng.ucsd.edu
[2] Telecom Paris, Institut Polytechnique de Paris, Palaiseau, France, 91120.
{han.qiu, gerard.memmi}@telecom-paris.fr
[3] Texas A&M University-Commerce, Texas, USA, 75428
meikang.qiu@tamuc.edu

Abstract. Deep Neural Networks (DNNs) in Computer Vision (CV) are well-known to be vulnerable to Adversarial Examples (AEs), namely imperceptible perturbations added maliciously to cause wrong classification results. Such variability has been a potential risk for systems in real-life equipped DNNs as core components. Numerous efforts have been put into research on how to protect DNN models from being tackled by AEs. However, no previous work can efficiently reduce the effects caused by novel adversarial attacks and be compatible with real-life constraints at the same time. In this paper, we focus on developing a lightweight defense method that can efficiently invalidate full whitebox adversarial attacks with the compatibility of real-life constraints. From basic affine transformations, we integrate three transformations with randomized coefficients that fine-tuned respecting the amount of change to the defended sample. Comparing to 4 state-of-art defense methods published in top-tier AI conferences in the past two years, our method demonstrates outstanding robustness and efficiency. It is worth highlighting that, our model can withstand advanced adaptive attack, namely BPDA with 50 rounds, and still helps the target model maintain an accuracy around 80 %, meanwhile constraining the attack success rate to almost zero.

Keywords: Adversarial Examples · Deep Learning · Security · Affine Transformation · Data Augmentation.

1 Introduction

With the rapid development of the Deep Neural Networks (DNNs) in Computer Vision (CV), there are more and more real-world applications that rely on the DNN models to classify images or to make decisions [12]. However, in recent years, the DNN models are well known to be vulnerable to Adversarial Examples (AE) which threats the robustness of the DNN usage [26]. Basically, the AEs can be

[1] Han Qiu is the corresponding author.

generated by adding carefully designed perturbations that are imperceptible to human eyes but can mislead DNN classifiers with very high accuracy [8].

Today, several rounds of AE attack and corresponding defense techniques have been developed as shown in [19]. The initial research on adversarial attacks on DNN models such as Fast Gradient Sign Method (FGSM) [8] aims at generating AEs by directly calculating the model gradients with respect to the input images. Such methods are then defeated by the defense methods based on various kinds of methods such as model distillation [17]. Then, the improved AE attacks are proposed to combine the gradient-based approach with the optimization algorithm such as the CW [4] aims to find the input features that made the most significant changes to the final output to mislead the DNN models. Such an optimized gradient-based approach can defeat many previous defense methods including the model distillation. Later, advanced defense methods are proposed to mitigate such attacks by obfuscating the gradients of the inference process. Specifically, some data augmentation techniques are deployed in such defense methods such as image compression, image denoising, image transformation [22], etc. Then, such state-of-the-art methods are then defeated by more advanced attack methods such as Backward Pass Differentiable Approximation (BPDA) [2] that can effectively approximate the obfuscated gradients to defeat these defenses.

In this paper, we propose a novel defense method that combines several data augmentation techniques together to mitigate the adversarial attacks against on the DNN models. We propose our method, Stochastic Affine Transformation (SAT), by deploying the image translation, image rotation, and image scaling method together. Our method can be used as a preprocessing step on the input images which makes our solution agnostic on many DNN models. Firstly, our method has little influence on the DNN inference which can effectively maintain the classification accuracy of benign images. Then, intensive experimentation and comparison have been performed to show the improvement of our method compared with several previous state-of-the-art defense solutions. Moreover, our method is a lightweight preprocess-only step that can be used on resource-constrained use cases such as the Internet of Things (IoT) [21].

This paper includes two main contributions. (1) We design a data augmentation-based defense solution to mitigate the initial and optimized gradient-based adversarial attacks on DNN models. Our method combining several steps of data augmentation techniques can be used as a preprocessing step on input images that can effectively maintain the agnostic DNN model's accuracy. (2) Our method can also defeat the advanced adversarial attack method such as BPDA which outperforms many previous state-of-the-art defense solutions.

This paper is organized as follows. Section 2 discusses the background information of this research including the brief definition of the adversarial examples and the previous data augmentation-based defense solutions. Section 3 presents our threat model and defense requirements. Section 4 proposes our methodology including the algorithm and the design details. Section 5 illustrates the experimentation details and evaluation results comparing with the previous state-of-the-art solutions. We then conclude in Section 6.

2 Research Backgrounds

In this section, we briefly introduce the background of the AEs in DNNs, the related work on AEs, and the state-of-the-art preprocessing-based defense methods based on data augmentation techniques.

2.1 Adversarial Examples in Deep Neural Networks

AEs can be explained as imperceptible modified samples that force one or multiple DNN models outputs with wrong results. This was first highlighted by [26]. By denoting I an input image, an adversarial example generated from it can be denoted as $\widetilde{I} = I + \delta$, where δ is the adversarial perturbation. The target model, which conducts inference for classification tasks can be denoted as f, thus the problem of performing adversarial attacks on the target DNN model can be formulated as Eq 1.

$$min\|\delta\|, \; s.t. \; f(\widetilde{I}) \neq f(I) \tag{1}$$

This equation can be interpreted as an optimization task that searches for a $\widetilde{I}$ based on I that can be misclassified by the target model, while keeping $\widetilde{I}$ visually as similar to I as possible. The aforementioned AE generation case is untargeted which aims to mislead the DNN classifier without a pre-set wrong label. As a targeted AE generation procedure aims to attack the DNN classifier to misclassify an input I with original label l as the pre-set wrong label l'. In the concern of real-life adoption of DNN models, both cases can result in serious outcomes if the models are not protected.

Since the time this vulnerability of DNNs has been discovered, various kinds of attacks have been proposed in the past few years to help the society to better understand the nature of AEs. To sum up, past work on adversarial attacks can be classified into two main approaches including initial gradient-based and optimized gradient-based. Fast Gradient Sign Method (FGSM) [8] is one of the most famous initial gradient-based adversarial attacks which calculates the model gradients based on the sign of the gradient of the classification loss concerning the input image. FGSM performs a one-step gradient update along the direction of the sign of gradient at each pixel under L_{inf} constraints to generate AEs. Later on, variations of FGSM were introduced to better searching for the optimum AE based on a single input. Such kind of methods includes I-FGSM [10] and MI-FGSM [6], aim at iteratively calculating the perturbations based on FGSM with a small step or with momentum.

Then, optimized gradient-based AE attacks are proposed to calculate the gradients based on adopting optimization algorithms to find optimal adversarial perturbations directly between the input images, and output predicted labels [4]. Such kind of attack is especially powerful in a whitebox or graybox scenario by adopting optimization algorithms to enhance the gradient calculation. Various optimized gradient-based AE attacks were proposed in recent years including Jacobian-based Saliency Map Attack (JSMA [16]), PGD [11], DeepFool [15], LBFGS [26], Carlini & Wagner (CW [4]), and Backward Pass Differentiable Approximation (BPDA [2]) technique. We should highlight the BPDA technique here, as it invalidates dozens of existing state-of-art defense approaches in recent evaluations [2]. The BPDA technique adopted in an attack that assumes that a defense function $g(\cdot)$ maintains the property $g(I) \approx I$ in order to preserve the functionality of the target model $f(\cdot)$. The adversary can then use $g(I)$ on the forward pass and replace it with I on the backward pass when calculating the gradients.

2.2 Data Augmentation based Preprocessing Defense Solutions

Various defensive strategies have been proposed to defeat adversarial attacks. One direction is to train a more robust model from either scratch or an existing model. Those approaches aim to rectify AEs' malicious features by including AEs into the training set [29], processing all the training data [33], or revising the DNN topology [17]. However, training a DNN model is very time and resource-consuming, especially for real-life cases, where models are more complicated. Besides, in real-life, DNN models are packed as closed-source applications and cannot be modified, thus those methods are not applicable. Most of all, the adversary can still adaptively generate AEs for the new models [2].

A more promising direction is to preprocess the input data to eliminate adversarial influence without touching the DNN model. These solutions are more suitable in the concern of real-life cases, as it is feasible, efficient, and lightweight. Thus, The preprocessing based defense is within the scope of this paper, as they do not require any laborious work with the DNN models, which made them competitive with most of the real-life defense scenarios. Below we describe some previous works and their limitations:

Feature Distillation (FD) [13] designed a compression method based on the JPEG compression but modified the quantization step. The basic idea is to measure the importance of input features for DNNs by leveraging the statistical frequency component analysis within the DCT of JPEG. It demonstrated a huge improvement in defending adversarial attacks compared with the standard JPEG compression method [20].

SHIELD [5] aims to randomize the quantization step by tuning the window size and quantization factors in the JPEG compression method. In SHIELD, the Stochastic Local Quantization (SLQ) method is used to divide an image into 8×8 blocks and applies a randomly selected JPEG compression quality (tuning quantization factors) to every block. The advantage is that the authors randomized the selective quantization steps which make the defense process different for different input images and make the adversarial attacks more difficult.

Bit-depth Reduction (BdR) [32] performs a simple type of quantization that can remove small (adversarial) variations in pixel values from an image. In the evaluation of that work, it demonstrates a more effective result comparing to adversarial training. However, recently developed attacks are not within the scope of that work, namely the BPDA-based attack.

Pixel Deflection (PD) [18] aims to add similar natural noises that are not sensitive to the DNN model. The idea of deflection is to randomly sample a pixel from an image and replace it with another randomly selected pixel from within a small square neighborhood. This could generate an artificial noise that affects little on the DNN model but can disturb the adversarial perturbations. Then, a BayesShrink denoising process is followed to recover the image content before this image is feed into the DNN model. The results of such a method are convincing since it introduces randomness into the preprocessing step and does not require any modification on the DNN model. However, the robustness of this method is significantly reduced if the attackers have knowledge of the preprocess step [2].

3 Threat Model and Defense Requirements

3.1 Threat Model

Untargeted attacks and targeted attacks are two major types of adversarial attacks. Untargeted attacks try to mislead the DNN models to an arbitrary label different from the correct one. On the other hand, targeted attacks only considering succeed when the DNN model predicts the input as one specific label desired by the adversary [4]. In this paper, we only evaluate the targeted attacks. The untargeted attacks can be mitigated in the same way.

We consider a full whitebox scenario, where the adversary has full knowledge of the DNN model and the defense method, including the network architecture, exact values of parameters, hyper-parameters, and the details of the defense method. However, we assume the random numbers generated in real-time are perfect with a large entropy such that the adversary cannot obtain or guess the correct values. Such a targeted full whitebox scenario represents the strongest adversaries, as a big number of existing state-of-the-art defenses are invalidated as shown in [28].

As for the adversary's capability, we assume the adversary is outside of the DNN classification system, and he is not able to compromise the inference computation or the DNN model parameters (e.g., via fault injection to cause bit-flips [23] or backdoor attacks [7]). What the adversary can do is to manipulate the input data with imperceptible perturbations. In the context of computer vision tasks, he can directly modify the input image pixel values within a certain range. We use l_∞ and l_2 distortion metrics to measure the scale of added perturbations: we only allow the generated AEs to have either a maximum l_∞ distance of $8/255$ or a maximum l_2 distance of 0.05 as proposed in [2].

3.2 Defense Requirements

Various life-concerned vital tasks are already implemented with DNN in real-life, e.g., video surveillance [27], face authentication [14], autonomous driving [30], network traffic identification [34], etc. Most of those cases' inference is conducted either locally in distributed computing units or remotely with the help of cloud servers. In the case where inference procedures are conducted locally, resources are highly constrained, thus only lightweight designs of defense can protect the system without draw extra burden over those units. Moreover, for both cases, previous resolutions, e.g. adversarial training or extra modifications over the target model would be considerably costly for real-life cases (where each sample is of large size, thus any kind of retraining can be laborious). As mentioned in the previous section, most of the existing defense methods either too complicated to be compatible with real-life constraints ([3,9,24]) or not capable of effectively reduce the impact brought by adversarial attacks ([18]).

To cope with those constraints in real-life adoption of DNNs, we believe the following properties should be taken considered when designing novel adversarial defense methods:

- Accuracy-preserving: they should not affect much on the prediction accuracy of the DNN model on clean data samples that processed by those methods.
- Security: they should be capable to effectively reduce effects brought by adversarial perturbations.
- Lightweight: the defense method should not be too heavy to impact the devices' or units' performance or operations, considering the limited onboard computing capabilities and resources.
- Generalization Ability: the defense method should neither require modifications over the target DNN's structure nor require any kind of retraining.

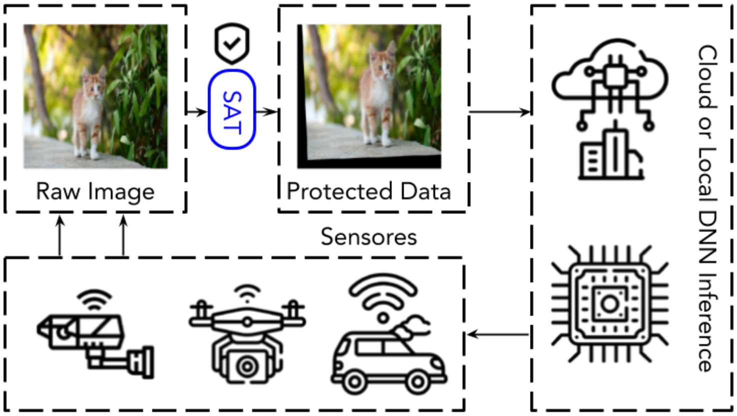


Fig. 1. A system overview of adopting SAT in the real-life use case.

Thus, for a better adaptation over nowadays real-life scenarios, we aim to design a preprocessing-only method to meet all those requirements. It has proved in our previous work [19], preprocessing-only adversarial defenses are competent enough to defense adversarial attacks, even for whitebox attacks.

4 Proposed Methodology

To better adapt to nowadays real-life scenarios' adoption of DNN, we present our efficient defense method against adversarial attacks that can conduct protections on the fly, termed Stochastic Affine Transformation (SAT). Thanks to the lightweight design of SAT, this defense should be more compatible with both cloud DNN inference as well as local or edge DNN inference respecting real-life scenarios. Fig. 1 illustrates an overview of adopting the SAT method in real life. The details of SAT will be present in Section 4.1. The analysis of the three hyperparameters respecting the defense efficiency is illustrated in Section 4.2.

4.1 SAT Algorithm

Following the logic of adding randomness to the affine transformation without harming the classification accuracy [19,9,31], we propose a simple but effective way of image distortion as an adversarial example defense. The algorithm is designed based on combining several affine transformation methods. The details are illustrated in Algorithm 1.

Three basic affine transformations with randomized coefficients are bounded tother in this single procedure, namely, translation, rotation, and scaling. We add randomness to those three simple affine transformations so that the attacker cannot utilize a useful gradient to generate adversarial examples even acknowledges the details of this defense. Such is done by acquiring different coefficients that follow three uniform distribution for different samples. To be specific, there are three coefficients along with the raw image as the input of the SAT method. T is the translation limit, R is the rotation limit, and S is the scaling limit. The original input will first be randomly shifted away from its original coordinates according to δ_x and δ_y that both follow the uniform distribution in the range $(-T, T)$. Then, the data will be randomly rotated at a certain angle δ_r that follows the uniform distribution in the range $(-R, R)$. Finally, the distorted image will be acquired by scaling up or down δ_s times, where δ_s follows a uniform distribution in the range $(1 - S, 1 + S)$.

Since only simple affine transformations and random number generator are adopted in SAT, we believe SAT is compatible with both cloud DNN inference procedures as well as localized or edge devices DNN inference procedure. This lightweight design is also hardware friendly and can conduct protections on the fly.

4.2 SAT Hyper-parameters

As aforementioned, there are three essential coefficients in the SAT method. In this part, we did a thorough evaluation and analysis of those three coefficients respecting the efficiency of defending adversarial attacks.

We believe a higher variance between the original data and the protected data while maintaining a high classification accuracy can help more to defend adversarial attacks, which is proved in our previous work [19]. In this part, three different metrics are adopted to evaluate this variance, namely the l_2 norm, Structural Similarity (SSIM) index, and the Peak Signal-to-Noise Ratio (PSNR). The classification accuracy (ACC) is the priority of most classification tasks thus is as well taken considerate in this part.

ALGORITHM 1: Stochastic Affine Transformation

```
Input: original image $I \in \mathbb{R}^{h \times w}$
Output: transformed image $I' \in \mathbb{R}^{h \times w}$
Parameters: translation limit $T$; scaling limit $S$, rotation limit $R$.
$I^{'} = O^{h \times w}$;
/* 1.Translation */
$\delta_x \sim \mathcal{U}(-T, T)$;
$\delta_y \sim \mathcal{U}(-T, T)$;
$\varDelta_x = \delta_x \times w$;
$\varDelta_y = \delta_y \times h$;
if $(x + \varDelta_x \in (0, w)) \wedge (y + \varDelta_y \in (0, h))$ then
    $I'(x, y) = I(x + \varDelta_x, y + \varDelta_y)$;
end
/* 2.Rotation */
$\delta_r \sim \mathcal{U}(-R, R)$;
$\varDelta_r = \delta_r \times \pi/180$;
for $(x_i, y_j)$ in $\{(x, y) | x \in (0, w), y \in (0, h)\}$ do
    $x_i^{'} = -(x_i - \lfloor w/2 \rfloor) \times sin(\varDelta_r) + (y_j - \lfloor h/2 \rfloor) \times cos(\varDelta_r)$;
    $y_j^{'} = (x_i - \lfloor w/2 \rfloor) \times cos(\varDelta_r) + (y_j - \lfloor h/2 \rfloor) \times sin(\varDelta_r)$;
    $x_i^{'} = \left\lfloor x_i^{'} + \lfloor w/2 \rfloor \right\rfloor$;
    $y_j^{'} = \left\lfloor y_j^{'} + \lfloor h/2 \rfloor \right\rfloor$;
    if $(x_i^{'} \in (0, w)) \wedge (y_j^{'} \in (0, h))$ then
        $I'(x_i, y_j) = I(x_i^{'}, y_j^{'})$;
    end
end
/* 3.Scaling */
$\delta_s \sim \mathcal{U}(1 - S, 1 + S)$;
$h_{new} = \delta_s \times h$;
$w_{new} = \delta_s \times w$;
$I' = \texttt{reshape}(I', (h_{new}, w_{new}))$;
if $\delta_s > 1$ then
    $I'(x, y) = cropping(I', (h, w))$;
end
if $\delta_s < 1$ then
    $I'(x, y) = padding(I', (h, w))$;
end
return $I'$;
```

l_2 is a widely adopted metric in deep learning domain to measure the amount of difference of two samples in the term of Euclidean distance, a higher l_2 indicates a greater difference. Eq. 2 shows how l_2 can be computed.

$$l_2(I', I) = \sqrt{(I'_R - I_R)^2 + (I'_G - I_G)^2 + (I'_B - I_B)^2}/(h \times w \times 3) \tag{2}$$

Where I' and I are the two 3-channel (RGB) samples to be compared. h and w are the height and width of those samples respectively.

SSIM is a metric in the computer vision domain normally being adopted to measure the similarity between two images, where smaller SSIM reflects greater difference. To be specific, SSIM is based on three comparison measurements between two samples, namely luminance (l), contrast (c), and structure (s). Each comparison function is elaborated in Eq 3, Eq 4, and Eq 5 respectively.

$$l(I', I) = \frac{2\mu_{I'}\mu_I + c_1}{\mu_{I'}^2 + \mu_I^2 + c_1} \tag{3}$$

$$c(I', I) = \frac{2\sigma_{I'}\sigma_I + c_2}{\sigma_{I'}^2 + \sigma_I^2 + c_2} \tag{4}$$

$$s(I', I) = \frac{\sigma_{I'I} + c_2/2}{\sigma_{I'}\sigma_I + c_2/2} \tag{5}$$

Where $\mu(\cdot)$ computes the mean of a sample, $\sigma^2(\cdot)$ computes the variance of a sample. $c1$ is equal to $(0.01 \times L)^2$, $c2$ is equal to $(0.03 \times L)^2$. Here L is the dynamic range of the pixel values. Finally, the SSIM can be acquired by computing the product of those three functions.

PSNR is most commonly used to measure the quality of reconstruction of lossy compression codecs, say compression, augmentation, or distortion, etc. A smaller PSNR indicates a greater difference between the two evaluating samples. PSNR can be defined via the mean squared error (MSE) between two comparing samples, which is explained in Eq 6.

$$PSNR(I', I) = 20 \cdot log_{10}(255) - 10 \cdot log_{10}(MSE(I', I)) \tag{6}$$

Where $MSE(\cdot)$ computes the MSE between two inputs.

We tried different values of T and S in the range $[0.01, 0.5]$. Different values of R is acquired in the range $[0, 40]$. Thus, each coefficient will test 11 values in the respecting range. As for different metrics, we will acquire 1331 ($11 \times 11 \times 11$) results from different combinations of those different values of coefficients. Fig. 2 demonstrates the change of those four metrics' value when different T, S,and R are adopted.

Table 1. The comparison of different methods over ACC and amount of changes.

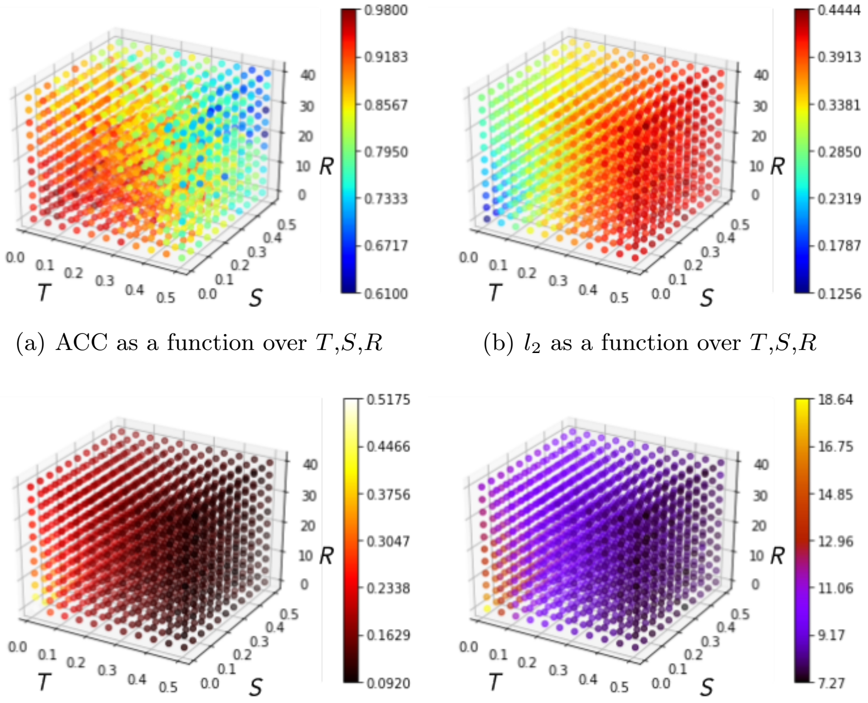
Defense	l_2 norm	SSIM	PSNR	ACC
SAT	**0.322**	**0.194**	**10.219**	**0.98**
FD [13]	0.1343	0.4310	18.050	0.97
SHIELD [5]	0.0405	0.8475	28.345	0.94
BdR [32]	0.0709	0.7730	23.010	0.92
PD [18]	0.0147	0.9877	37.100	0.97

In Fig. 2(a), the changes of ACC with those three coefficients varies is presented. The right side of Fig. 2(a) is the color-bar that reflecting the ACC attained with respecting combinations of those three coefficients. We can learn that lower T, S, and R can help the model maintain a high ACC. To ensure a high ACC, we set 95% ACC as a standard, thus those combinations with ACC below this standard would not be taken further considerations.

From Fig. 2(b) we can learn that the l_2 is not that sensitive with S and R comparing to T in their respecting range. Combining the information provided from Fig. 2(b), Fig. 2(c), and Fig. 2(d) we can as well acquire this similar analytical result for SSIM and PSNR. This phenomenon also reflects that the l_2, SSIM, and PSNR can all reflecting the scale of changes in a similar manner respecting affine transformations.

By overlapping Fig. 2(a) with the other three figures, we can acquire a set of optimum coefficients that ensures high ACC and great variance at the same time. The set of coefficients for the following experiment is set as follows: $T = 0.16$, $S = 0.16$, and finally $R = 4$.

We compare the SAT method using those fine-tuned hyperparameters with other state-of-art adversarial defense methods, which presented in Table 1. As demonstrated, SAT can create a greater difference between raw samples and protected samples while maintaining a high ACC than other methods. We will evaluate whether this greater variance will help and how much will it help DNN models to defend adversarial attacks in the following section.



(a) ACC as a function over T,S,R

(b) l_2 as a function over T,S,R

(c) SSIM as a function over T,S,R

(d) PSNR as a function over T,S,R

Fig. 2. Metrics of reconstructed images under different values of T,S,R

5 Experimentation and Evaluation

In this section, we conduct a comprehensive evaluation of the proposed technique. Various adversarial attacks are taken considered in this part: 4 kinds of standard adversarial attacks (FGSM, I-FGSM, LBFGS, and C&W) are conducted, advanced interactive gradient approximation attack, namely BPDA, is also conducted to evaluate the robustness of SAT. We compare SAT with four state-of-art defense methods publish in top-tier artificial intelligence conferences from the past two years. This section would be divided into three parts to elaborate on the settings of the experiment, efficiency over standard adversarial attacks, and the efficiency over BPDA respectively.

5.1 Experimental Settings

Tensorflow [1] is adopted as the deep learning framework to implement the attacks and defenses. The learning rate of the C&W and BPDA-based PGD attack is set to 0.1. All the experiments were conducted on a server equipped with 8 Intel I7-7700k CPUs and 4 NVIDIA GeForce GTX 1080 Ti GPU.

SAT is of general-purpose and can be applied to various models over various platforms as a preprocessing step for computer vision tasks as illustrated in Fig. 1. Without the loss of generality, we choose a pre-trained Inception V3 model [25] over the ImageNet dataset as the target model. This state-of-the-art model can reach 78.0% top-1 and 93.9% top-5 accuracy. We randomly select 100 images from the ImageNet Validation dataset for AE generation. These images can be predicted correctly by this Inception V3 model.

We consider the targeted attacks where each target label different from the correct one is randomly generated [2]. For each different attack, we measure the classification accuracy of the generated AEs (ACC) and the attack success rate (ASR) of the targeted attack. To be noticed that the untargeted attacks are not within our scope in this work. A higher ACC or lower ASR indicates the defense is more resilient against the attacks.

For comparison, we re-implemented 4 existing solutions including FD [13], SHIELD [5], Bit-depth Reduction [32], and PD [18].

5.2 Evaluation on Defending Adversarial Attacks

We first evaluate the efficiency of our proposed method over standard adversarial attacks, namely FGSM, I-FGSM, C&W, and LBFGS. For FGSM and I-FGSM, AEs are generated under l_∞ constraint of 0.03. For LBFGS and C&W, the attack process is iterated under l_2 constraint and stops when all targeted AEs are found. We measure the model accuracy (ACC) and attack success rate (ASR) with the protection of SAT and other defense methods.

The results are shown in Table 2 and Table 3. For benign samples only, our proposed techniques have the smallest influence on the model accuracy comparing to past works. For defeating AEs generated by these standard attacks, the attack success rate can be kept around 0% and the model accuracy can be drastically

Table 2. Comparisons of different defense against attacks respecting ACC.

Defense	**Clean**	**FGSM(ϵ=.03)**	**IFGSM(ϵ=.03)**	**C&W**	**LBFGS**
Baseline	1.00	0.42	0.02	0.06	0.00
SAT	**0.98**	**0.61**	**0.85**	**0.78**	**0.96**
FD [13]	0.97	0.47	**0.87**	**0.84**	**0.97**
SHIELD [5]	0.94	0.49	0.84	**0.78**	0.92
BdR [32]	0.92	0.47	0.82	0.61	0.90
PD [18]	0.97	0.42	0.30	0.11	0.86

recovered, which an efficiency against different kinds of adversarial attacks is demonstrated. To be noticed, previous work can only attain an accuracy of around 50% on samples attacked by the FGSM ($\epsilon = .03$). SAT can recover the accuracy to 0.61%.

Comparing to the Table 1 which compares different methods' capability of creating a variance between input and defended sample, the defense efficiency evaluated in this part shows a strong correlation with the amount of variance generated by the defense method. This has confirmed the previous conclusion in our previous work [19].

In a nutshell, the effectiveness of SAT against standard adversarial attack is demonstrated, as we can attain a state-of-art defense efficiency on all the evaluated attacks comparing to other methods.

Table 3. Comparisons of different defense against attacks respecting ASR.

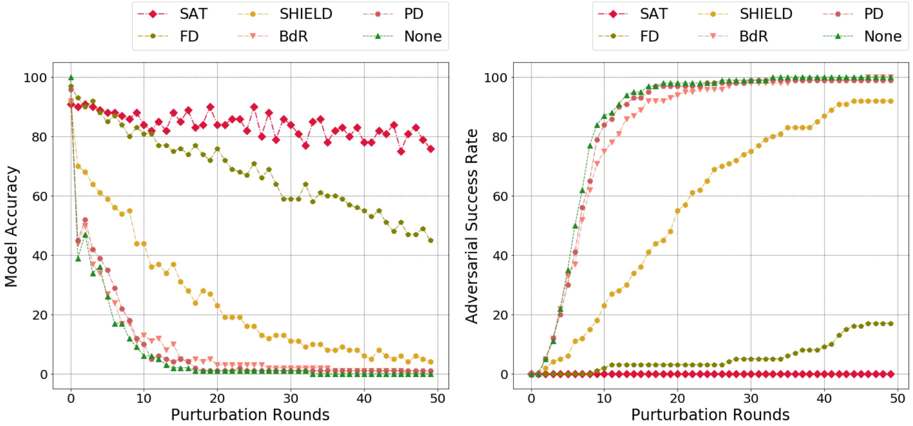
Defense	**IFGSM(ϵ=.03)**	**C&W**	**LBFGS**
Baseline	0.95	1.00	1.00
SAT	**0.01**	**0.01**	**0.00**
FD [13]	**0.00**	**0.06**	**0.00**
SHIELD [5]	**0.01**	0.02	**0.00**
BdR [32]	0.02	0.2	**0.00**
PD [18]	0.6	0.11	0.02

5.3 Evaluation on Defending Advanced Adversarial Attacks

We then evaluate the effectiveness of SAT against the BPDA-based PGD attack. Since the BPDA-based PGD attack is an interactive attack, we record the ACC and ASR for each round for different defense methods.

The model prediction accuracy and attack success rate in each round are shown in Fig. 3(a) and 3(b), respectively. We can observe that after 50 attack rounds, all other three prior solutions except FD can only keep the model accuracy

[2] The ASR evaluation of the FGSM is not shown here since the baseline ASR is 0.



(a) ACC per round under BPDA-based attack.

(b) ASR per round under BPDA-based attack.

Fig. 3. ACC and ASR of various techniques under the BPDA-based attack.

lower than 5%, and attack success rates reach higher than 90%. Those defenses fail to mitigate the BPDA-based attack. FD can keep the attack success rate lower than 20% and the model accuracy is around 40%. This is better but still not very effective in maintaining the DNN model's robustness.

In contrast, SAT is particularly effective against the BPDA-based attack. As our method can maintain an acceptable model accuracy (around 80% for 50 perturbation rounds), and restrict the attack success rate to 0 for all the record rounds. This result is as well consistent with the l_2, SSIM, and PSNR metrics compared in Table 1: the randomization effects in SAT cause greater variances between I' and I, thus invalidating the BPDA-based attack basic assumption, which is $I' \approx I$.

To sum up, the effectiveness of SAT against the BPDA-based attack is demonstrated, as a considerable improvement over the defense efficiency against the BPDA-based attack is shown comparing to previous work.

6 Conclusion

In this paper, we proposed a lightweight defense method that can effectively invalidate adversarial attacks, termed SAT. By adding randomness to the coefficients, we integrated three basic affine transformations into SAT. Compared with four state-of-art defense methods published in the past two years, our method clearly demonstrated a more robust and effective defense result on standard adversarial attacks. Moreover, respecting the advanced BPDA-based attack, SAT showed an outstanding capability of maintaining the target model's ACC and detain the ASR to 0. This result is almost 50% better than the best result achieved by previous work against full whitebox targeted attacks.

References

1. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al.: Tensorflow: A system for large-scale machine learning. In: 12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16). pp. 265–283 (2016)
2. Athalye, A., Carlini, N., Wagner, D.: Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In: International Conference on Machine Learning. pp. 274–283 (2018)
3. Buckman, J., Roy, A., Raffel, C., Goodfellow, I.: Thermometer encoding: One hot way to resist adversarial examples (2018)
4. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy (SP). pp. 39–57. IEEE (2017)
5. Das, N., Shanbhogue, M., Chen, S.T., Hohman, F., Li, S., Chen, L., Kounavis, M.E., Chau, D.H.: Shield: Fast, practical defense and vaccination for deep learning using jpeg compression. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 196–204 (2018)
6. Dong, Y., Liao, F., Pang, T., Hu, X., Zhu, J.: Discovering adversarial examples with momentum. arXiv preprint arXiv:1710.06081 (2017)
7. Gao, Y., Xu, C., Wang, D., Chen, S., Ranasinghe, D.C., Nepal, S.: Strip: A defence against trojan attacks on deep neural networks. In: Proceedings of the 35th Annual Computer Security Applications Conference. pp. 113–125 (2019)
8. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
9. Guo, C., Rana, M., Cisse, M., van der Maaten, L.: Countering adversarial images using input transformations. In: International Conference on Learning Representations (2018)
10. Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial examples in the physical world. arXiv preprint arXiv:1607.02533 (2016)
11. Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial machine learning at scale. arXiv preprint arXiv:1611.01236 (2016)
12. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature **521**(7553), 436–444 (2015)
13. Liu, Z., Liu, Q., Liu, T., Xu, N., Lin, X., Wang, Y., Wen, W.: Feature distillation: DNN-oriented jpeg compression against adversarial examples. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 860–868. IEEE (2019)
14. Mao, Y., Yi, S., Li, Q., Feng, J., Xu, F., Zhong, S.: A privacy-preserving deep learning approach for face recognition with edge computing. In: Proc. USENIX Workshop Hot Topics Edge Comput.(HotEdge). pp. 1–6 (2018)
15. Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P.: Deepfool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2574–2582 (2016)
16. Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z.B., Swami, A.: The limitations of deep learning in adversarial settings. In: 2016 IEEE European symposium on security and privacy (EuroS&P). pp. 372–387. IEEE (2016)
17. Papernot, N., McDaniel, P., Wu, X., Jha, S., Swami, A.: Distillation as a defense to adversarial perturbations against deep neural networks. In: 2016 IEEE Symposium on Security and Privacy (SP). pp. 582–597. IEEE (2016)

18. Prakash, A., Moran, N., Garber, S., DiLillo, A., Storer, J.: Deflecting adversarial attacks with pixel deflection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8571–8580 (2018)
19. Qiu, H., Zeng, Y., Zheng, Q., Zhang, T., Qiu, M., Memmi, G.: Mitigating advanced adversarial attacks with more advanced gradient obfuscation techniques. arXiv preprint arXiv:2005.13712 (2020)
20. Qiu, H., Zheng, Q., Memmi, G., Lu, J., Qiu, M., Thuraisingham, B.: Deep residual learning based enhanced jpeg compression in the internet of things. IEEE Transactions on Industrial Informatics (2020)
21. Qiu, H., Zheng, Q., Zhang, T., Qiu, M., Memmi, G., Lu, J.: Towards secure and efficient deep learning inference in dependable iot systems. IEEE Internet of Things Journal (2020)
22. Qiu, M., Qiu, H.: Review on image processing based adversarial example defenses in computer vision. In: 2020 IEEE 6th Intl Conference on Big Data Security on Cloud (BigDataSecurity). pp. 94–99. IEEE (2020)
23. Rakin, A.S., He, Z., Fan, D.: Bit-flip attack: Crushing neural network with progressive bit search. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1211–1220 (2019)
24. Shaham, U., Yamada, Y., Negahban, S.: Understanding adversarial training: Increasing local stability of supervised models through robust optimization. Neurocomputing **307**, 195–204 (2018)
25. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016)
26. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
27. Tang, Y., Zhang, C., Gu, R., Li, P., Yang, B.: Vehicle detection and recognition for intelligent traffic surveillance system. Multimedia tools and applications **76**(4), 5817–5832 (2017)
28. Tramer, F., Carlini, N., Brendel, W., Madry, A.: On adaptive attacks to adversarial example defenses. arXiv preprint arXiv:2002.08347 (2020)
29. Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., McDaniel, P.: Ensemble adversarial training: Attacks and defenses. arXiv preprint arXiv:1705.07204 (2017)
30. Wu, B., Iandola, F., Jin, P.H., Keutzer, K.: Squeezedet: Unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 129–137 (2017)
31. Xie, C., Wang, J., Zhang, Z., Ren, Z., Yuille, A.: Mitigating adversarial effects through randomization. In: International Conference on Learning Representations (2018)
32. Xu, W., Evans, D., Qi, Y.: Feature squeezing: Detecting adversarial examples in deep neural networks. In: 25th Annual Network and Distributed System Security Symposium, NDSS 2018, San Diego, California, USA, February 18-21, 2018. The Internet Society (2018)
33. Yang, Y., Zhang, G., Katabi, D., Xu, Z.: Me-net: Towards effective adversarial robustness with matrix estimation. In: International Conference on Machine Learning. pp. 7025–7034 (2019)

34. Zeng, Y., Gu, H., Wei, W., Guo, Y.: $deep-full-range$: A deep learning based network encrypted traffic classification and intrusion detection framework. IEEE Access **7**, 45182–45190 (2019)