ECAI 2020
G.D. Giacomo et al. (Eds.)

doi:10.3233/FAIA200247

Structure Matters: Towards Generating Transferable Adversarial Images

Dan Peng[1] and **Zizhan Zheng** [2] and **Linhao Luo**[1] and **Xiaofeng Zhang**[1]

Abstract. Recent works on adversarial examples for image classification focus on directly modifying pixels with minor perturbations. The small perturbation requirement is imposed to ensure the generated adversarial examples being natural and realistic to humans, which, however, puts a curb on the attack space thus limiting the attack ability and transferability especially for systems protected by a defense mechanism. In this paper, we propose the novel concepts of structure patterns and structure-aware perturbations that relax the small perturbation constraint while still keeping images natural. The key idea of our approach is to allow perceptible deviation in adversarial examples while keeping structure patterns that are central to a human classifier. Built upon these concepts, we propose a *structure-preserving attack (SPA)* for generating natural adversarial examples with extremely high transferability. Empirical results on the MNIST and the CIFAR10 datasets show that SPA exhibits strong attack ability in both the white-box and black-box setting even defenses are applied. Moreover, with the integration of PGD or CW attack, its attack ability escalates sharply under the white-box setting, without losing the outstanding transferability inherited from SPA.

1 INTRODUCTION

Deep neural networks (DNNs) have achieved phenomenal success in computer vision by showing superior accuracy over traditional machine learning algorithms. However, recent works have demonstrated that DNNs are vulnerable to adversarial examples that are generated for malicious purposes [35, 16].This observation has raised serious concerns on the robustness of the state-of-the-art DNNs and limited their applications in various security-sensitive applications [15, 34].

Generally speaking, adversarial examples can be any valid inputs to machine learning models that are intentionally designed to cause mistakes [14]. Although intuitively, an input is valid as long as it is natural and meaningful to human eyes, how to quantify this formally in the objective function is challenging. For object recognition, we rely on human labelers to obtain the ground truth labels, which are unknown to the attacker before natural adversarial examples are generated. To bypass this dilemma, a common attack strategy is to start with a clean image where the ground truth label is already known and modify it so that the new image is natural and semantically similar to the original image while its output label differs from the ground truth label of the clean image.

A simple approach for obtaining natural adversarial examples and ensuring semantic similarity that has been intensively studied in the literature is to introduce small perturbations into pixels such that the distortion between the adversarial example and the original image is humanly imperceptible [16, 33, 10, 9]. Note that the perturbation considered in these works is typically unstructured as random noise, thus only very small perturbations can be allowed to be superimposed onto images; Otherwise, the large unstructured distortion will destroy the semantics of the original image and further make the generated image unnatural and less meaningful. We highlight that the small perturbation requirement is not a necessity. The ultimate goal is to ensure the generated images being both natural and realistic to human eyes.

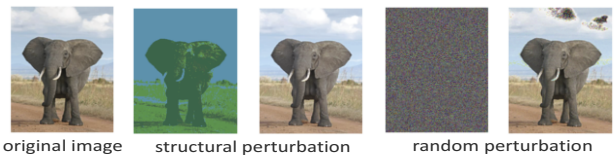

Figure 1: An example of structural perturbations. The second and fourth images are the structural and random perturbations, respectively, under the same maximum allowed distortion size. The two images are scaled (with pixel values enlarged by a factor of 10) for clear illustration. The third and fifth images are the adversarial images generated from the original image with the corresponding perturbations added. Structural adversarial examples are more natural than adversarial examples generated by adding random perturbations.

However, the small perturbation size leads to the low transferability of these perturbation-based attacks (see the detailed analysis in Section 4.3). Thus, there is a demand for new approaches that can tolerate larger distortion while still ensuring the generated adversarial examples being natural and sharing the same semantics of original images.

In this paper, we propose a *structure-preserving attack (SPA)* for generating natural adversarial examples with high transferability. Our approach is based on the hypothesis that the semantics of an image is mainly derived from its spatial structures [22]. Thus, a promising approach to ensure semantic similarity is to maintain the core structural patterns across images. The main idea of our approach is to introduce *structural perturbations* to images so that the generated adversarial images keep similar structures as the original images. Instead of giving an accurate definition of *structures*, which is a challenging task, we adopt an intuitive definition of *structure pattern* by partitioning pixels according to their intensity values (See Definition 1). In our SPA algorithm, we enforce that the same perturbation is applied to all the pixels in the same structure pattern so that the computed perturbation for the given image is structural. By imposing the structural constraints, our approach can tolerate moderate to large distortion while still guaranteeing that the generated images are realistic and

[1] Harbin Institute of Technology (Shenzhen), China, email: {pengdan,luolinhao}@stu.hit.edu.cn, zhangxiaofeng@hit.edu.cn. Xiaofeng Zhang is the corresponding author.
[2] Tulane University, USA, email: zzheng3@tulane.edu.

meaningful to human eyes. Consequently, SPA allows relatively large distortion to the original images than typical small-perturbation-based attacks, leading to better attack ability and transferability. This idea is further illustrated in Figure 1, where structural and random perturbations are added to the original image, respectively, with the same maximum allowed distortion size of 20/255. We observe that unlike random perturbation, the structural perturbation maintains the structure of the original image and is semantically meaningful to human beings. Thus, even under relatively larger distortion, the adversarial example generated through structural perturbation is still natural. In contrast, the random perturbation may destroy the structure of the original image and further dirty the image to some extent.

We show that our structure-preserving adversarial examples are highly transferable with little loss of successful attack rate when applied to black-box attacks even when a defense mechanism [25, 13] is applied.

This work broadens the scope of adversarial machine learning by showing a new class of adversarial examples that follow different distributions from the training dataset while still being legible and natural to humans. Our study reveals the weakness of current defense mechanisms in the face of structure-preserving attacks that relax the small perturbation constraint.

Our main contributions can be summarized as follows.

- We propose the structure-preserving attack (SPA) as a new approach for generating natural adversarial examples with strong transferability. The proposed structural perturbation concept is a general idea and can be combined with small-perturbation-based attacks to generate even stronger attacks.
- We conduct comprehensive experiments and demonstrate that SPA adversarial examples achieve extremely high transferability even when defense mechanisms are applied. Further, when combined with other attacks, SPA strikingly enhances both the white-box and black-box attack abilities.
- We conduct adversarial training with SPA and show that even SPA-based adversarial training hardly resist SPA itself, which further demonstrates the effectiveness of SPA.
- We analyze the relationship between attack ability and attack space from the perspective of *space flexibility* and *distortion flexibility*. We show that to obtain strong attack ability, it is profitable to sacrifice a bit of space flexibility in exchange for greater distortion flexibility.

2 BACKGROUND

In this section, we briefly review recent studies on adversarial examples and defense mechanisms.

2.1 Adversarial attacks

Many adversarial attacks have been proposed [9, 16, 25, 10], most of which are small-perturbation-based (measured by a L_p-norm for some p). Among them, the Projected Gradient Descent (PGD) method is the most effective L_∞-norm based attack for naturally trained networks and has good transferability, while L_2-norm based CW attack is the most effective white-box attack for deterministic networks, including both naturally trained networks and PGD-based adversarially trained networks. A neural network is considered deterministic if the model does not utilize any randomization.

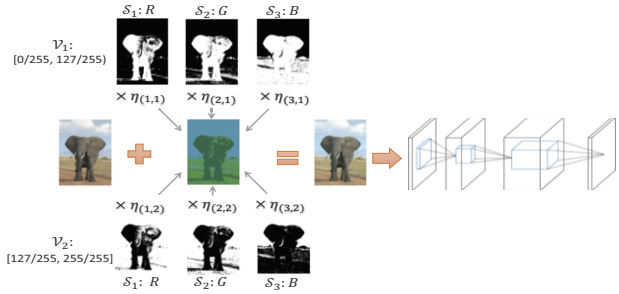

Figure 2: The overall network architecture for generating SPA adversarial examples. The SPA layer sits in front of the target model.

2.2 Defense techniques

To mitigate the threat of adversarial examples, a number of defense mechanisms have been proposed in the literature [21, 25, 17, 12]. However, most defense methods are ineffective in the newly proposed attacks [3, 8]. Only a few state-of-the-art defense models have demonstrated their robustness to adversarial examples [25, 13]. In particular, PGD-based adversarial training [25] is an effective defense to resist L_∞ attacks include PGD itself, while randomization has been shown to be an effective technique to resist the L_2-based CW attack as the randomness in target networks makes it difficult to compute accurate adversarial perturbations. A representative randomization technique is randomized smoothing, where a Gaussian noise layer is sited in the front of the target model [13]. To shed light on the weakness of existing defense mechanisms, we will evaluate SPA against PGD-based adversarial training and randomized smoothing. Details are given in the evaluation section.

3 STRUCTURE-PRESERVING ATTACKS TO DNNS

The proposed approach is called **S**tructure-**P**reserving **A**ttack (SPA), which attempts to generate structural adversarial perturbations that maintain the same structure as the original images. Similar to previous small-perturbation-based attacks, our structural adversarial examples are generated by introducing perturbations to the original images. The key difference, however, is that instead of perturbing each pixel in an image independently as in most previous work, pixels [3] in the same structure pattern (to be formally defined below) are perturbed similarly in our approach. By considering the intrinsic structure of images, structural adversarial examples generated under relatively large perturbations are comparably natural or even more natural than traditional small-perturbation-based adversarial examples (see our experiment results in the next section). Thus, SPA exhibits stronger transferability.

SPA can be either untargeted or targeted and can be used for both white-box and black-box attacks. In this paper, we focus on untargeted white-box and black-box attacks. More specifically, for a given image x, we attempt to find a small perturbation ϵ so that the adversarial perturbation ϵ keeps the structure of the original image, while the prediction on new image $x' = x + \epsilon$ reported by the target model f is different from the ground truth label $C_{\text{true}}(x)$. Note that $C_{\text{true}}(x)$ is the ground truth of the original image x. However, since the adversarial example should be semantically similar to the original image,

[3] By definition, a pixel in an image is the composition of all the channels in a specific position in the image. In this work, a pixel sometimes refers to the feature of a single channel in a specific position when there is no ambiguity.

they share the same ground truth. Formally, we require that

$$\arg\max_{y\in\mathcal{Y}} f_y(x') \neq C_{\text{true}}(x), \quad \|\epsilon\|_p \leq \delta, \text{ and } \quad \epsilon \sim x, \tag{1}$$

where we use '$\sim$' to denote the fact that the structures of two images are similar to each other. Notice that the first two conditions are commonly adopted in small-perturbation-based adversarial examples, while the last condition is unique to our approach. By explicitly introducing the structure-preserving perturbation constraint, our approach can tolerate relatively large perturbations while still keeping the generated adversarial examples natural. The key idea that differentiates our approach from typical small-perturbation-based attacks is to ensure the perturbations being structural, which in turn implies structural adversarial examples. Although small-perturbation-based adversarial examples also keep the structure of original images, this is completely achieved through the small perturbation restriction. In contrast, structure-preserving is achieved through the structural perturbation restriction in our approach, which is orthogonal to the small perturbation restriction. This is the key reason why our approach may relax the small perturbation restriction.

We then discuss our approach for measuring structure similarity. Intuitively, a structure in an image refers to a continuous region of pixels that constitute an object or the background aligned with human perception. For example, we may consider the elephant, cloud, blue sky, road and meadow as structure patterns in Figure 1. Different structures (objects) can be roughly identified based on pixel intensities as a structure (object) in an image often consists of pixels of similar colors. Therefore, although it is difficult to give an accurate definition of structures, we approximate a structure pattern with all pixels with similar pixel values, which is formally defined as follows:

Definition 1 ***Structure Pattern:*** *Let $\mathcal{S}$ be a space partition that divides the set $[1,...,W]\times[1,...,H]\times[1,...,C]$ into disjoint sub-regions. For the s-th sub-region in $\mathcal{S}$, let $\mathcal{V}_s$ denote a partition that divides the interval $[0,1]$ into disjoint sub-intervals. A structure pattern of indices (s,v) is defined as the set of pixels for which the pixel position lies in the s-th sub-region in $\mathcal{S}$ and the pixel value lies in the v-th sub-interval in $\mathcal{V}_s$. For simplicity, we assume $|\mathcal{V}_s|$ is the same for all s, and let $S=|\mathcal{S}|$ and $V=|\mathcal{V}_s|$.*

According to this definition, a structure pattern consists of a set of pixels that share similar pixel values and are in close proximity. It is worth noting that in traditional L_p-norm based perturbations, each pixel is treated *independently*, which can be viewed as a specific case of our definition, where each structure pattern has a single pixel. Note that Definition 1 provides a general definition of structure patterns and applies to any possible partitions $\mathcal{S}$ and $\{\mathcal{V}_s\}$. In our experiment discussed below, the whole space is partitioned into three sub-regions corresponding to the three channels (i.e., $S=3$). Further, an even partition of pixel values is applied to all the channels. Figure 2 gives an example of this approach where $S=3$ and $V=2$. This simple approach for generating partitions already provides satisfactory results. However, our approach discussed below applies to any given partitions.

Our main idea for keeping structure patterns unchanged under perturbations is to guarantee that all the pixels in the same structure pattern are perturbed by a *similar* amount. That is, instead of perturbing each pixel independently as in most previous work, we consider structural perturbations that are aligned with the structure patterns of images. To simplify the implementation, we consider a special case in this work by requiring that all the pixels in the same structure pattern are perturbed by the *same* amount.

Formally, we define a *meta-perturbation* as an $S\times V$ matrix η where $\eta(s,v)$ gives the perturbation to be added to the pixels in the structure pattern of indices (s,v). Let $\epsilon=\texttt{pert}(x,\eta)$, where $\texttt{pert}$ is a function that generates the perturbation ϵ for image x according to the meta-perturbation η. An implementation of $\texttt{pert}$ is given in Algorithm 1 (lines 9-19), where for each pixel, its structure indices in the space and value partitions are first identified and the corresponding meta-perturbation value is then obtained. Algorithm 1 can be made more efficient through vectorization. We choose the current form to illustrate the main idea more clearly. Alternatively, for each structure pattern of indices (s,v) in an image x, we may define a binary mask (a black-white image of the same shape as image x) $b_{s,v}$, where $b_{s,v}(m,n,c)=1$ if the pixel (m,n,c) is in the structure pattern (s,v) and $b_{s,v}(m,n,c)=0$ otherwise. Figure 2 shows the six binary masks for the original image x on the left. The desired perturbation for x can then be found by taking a weighted sum of the binary masks with the weights taken from η. That is, $\texttt{pert}(x,\eta)=\sum_{s,v}\eta(s,v)b_{s,v}$.

Our objective is to find a meta-perturbation η that meets the following condition:

$$\arg\max_{y\in\mathcal{Y}} f_y(x+\texttt{pert}(x,\eta)) \neq C_{\text{true}}(x), \quad \|\texttt{pert}(x,\eta)\|_p \leq \delta \tag{2}$$

It is important to mention that in addition to the misclassification goal and the small perturbation restriction, we further require that the generated perturbation preserves the structure patterns of the original images, which is achieved by the $\texttt{pert}$ function. Also note that $\|\texttt{pert}(x,\eta)\|_p=\|\eta\|_p$ for $p=\infty$.

3.1 Generating structural adversarial perturbations

SPA is trained together with the target model by adding an extra layer in front of the target network (see Figure 2). For untargeted attacks, our objective is to find a small meta-perturbation η so that $\arg\max_{y\in\mathcal{Y}} f_y(x+\texttt{pert}(\eta,x)) \neq C_{\text{true}}(x)$ for any $x\in\mathcal{X}$. To improve the chance of successful attacks, we aim to find η so that the distance between the predicted logits and the ground truth is maximized for a given image. Let $l(x)$ be a $|\mathcal{Y}|$ dimensional vector where $l_y(x)=1$ if $y=C_{\text{true}}(x)$ and $l_y(x)=0$ otherwise. We then solve the following optimization problem to find the meta-perturbation η for a given image x:

$$\begin{aligned}\operatorname*{argmax}_{\eta} \ & \mathcal{J}(f(x+\texttt{pert}(x,\eta)),l(x)) \\ s.t. \ & \|\texttt{pert}(x,\eta)\|_p \leq \delta\end{aligned} \tag{3}$$

where $\mathcal{J}$ is a loss function that measures the difference between the output logit $f(x+\texttt{pert}(\eta,x))$ when the target model f is applied to the crafted adversarial example $x+\texttt{pert}(\eta,x)$ and the ground truth of the original image. In this work, we use the cross entropy as the loss function.

As SPA sits in front of the target network as shown in Figure 2, we fix the target model when solving the optimization problem (3) to generate SPA adversarial examples. Projected gradient descent (PGD) is a standard technique to solve L_p-constrained optimization problems [5]. It has recently been used to design adversarial attacks and PGD attack has become a benchmark attack [25] [4]. In this paper, we use PGD to solve the above constrained optimization problem for finding the optimal parameters η and adopt L_∞ as the norm metric as

[4] In this work, PGD refers to PGD attack unless otherwise specified.

Algorithm 1: Structure-Preserving Attack against DNNs

Input: target model f; original labelled images $(x^{[W,H,C]}, l(x))$; a space partition $\mathcal{S}$ and a set of pixel value partitions $\{\mathcal{V}_s\}$ where $|\mathcal{S}| = S$ and $|\mathcal{V}_s| = V$ for any s; maximum perturbation size δ; step size σ; number of steps K

Output: meta-perturbation η

1 $\eta^{(1)} = 0^{[S,V]}$;
2 **for** $k \in [1, ..., K]$ **do**
3 $\quad \epsilon = \texttt{pert}(x, \eta^{(k)})$;
4 $\quad L = \mathcal{J}(f(x+\epsilon), l(x))$;
5 $\quad \bigtriangledown_{\eta^{(k)}} L = \frac{\partial L}{\partial \eta} |_{\eta^{(k)}}$;
6 $\quad \eta^{(k+1)} = \eta^{(k)} + \sigma \times sign(\bigtriangledown_{\eta^{(k)}} L)$;
$\quad$ // project each entry in η into $[-\delta, \delta]$
7 $\quad \eta^{(k+1)} = \texttt{clip}(\eta^{(k+1)}, -\delta, \delta)$;
8 **end**
9 **def** pert (x, η) **:**
10 $\quad$ **for** $c \in [1, ..., C]$ **do**
11 $\quad\quad$ **for** $m \in [1, ..., W]$ **do**
12 $\quad\quad\quad$ **for** $n \in [1, ..., H]$ **do**
$\quad\quad\quad\quad$ // find the sub-region in $\mathcal{S}$ where the pixel is in
13 $\quad\quad\quad\quad s = \texttt{space\_index}(m, n, c, \mathcal{S})$;
$\quad\quad\quad\quad$ // find the sub-interval in $\mathcal{V}_s$ where the pixel is in
14 $\quad\quad\quad\quad v = \texttt{value\_index}(x(m, n, c), \mathcal{V}_s)$;
15 $\quad\quad\quad\quad \epsilon(m, n, c) = \eta(s, v)$;
16 $\quad\quad\quad$ **end**
17 $\quad\quad$ **end**
18 $\quad$ **end**
19 $\quad$ **return** ϵ

in PGD attack. We highlight that our SPA approach is a general idea and can be applied to most existing attacks [16, 9, 10] [5].

The sign value of the gradient multiplied by a constant step size σ is then used to update η (line 6). We then project η to satisfy the small perturbation constraint. For L_∞-norm based perturbation constraint, this can be easily implemented using the clip function to restrict the perturbation to fall into the maximum allowed distortion range $[-\delta, \delta]$ (line 7). We highlight that instead of computing the gradient with respect to the perturbation itself as in standard PGD attack [25], we compute the gradient with respect to the meta-perturbation and then use the computed meta-perturbation to form the structural perturbation. This is crucial for ensuring the perturbation generated by SPA being structural and is the main difference between SPA and PGD.

Due to the good convergence property of the cross entropy loss function and the small number of parameters ($S \times V$ parameters) to be optimized, the searching for the optimal meta-perturbation η converges within a small number of iterations.

4 EXPERIMENT RESULTS

To evaluate the performance of SPA, we compare it with two baseline attack algorithms, PGD attack [25] and CW attack [9] on two popular image classification datasets, MNIST [23], and CIFAR10 [20]. Moreover, the attack methods are evaluated both on the vanilla models with different architectures and when the two state-of-the-art defense mechanisms, PGD-based adversarial training [25] and randomized smoothing [13], are applied.

4.1 Experiment settings

Evaluation metrics: As in most previous works, we report the classification accuracy of target models under various attack-defense configurations. For L_2-based CW attack, both the accuracy and distortion size determine its attack ability. Thus, we also report the L_2 distortion size for CW attack.

A stronger attack method leads to a lower classification accuracy while a target model with stronger defense ability has a higher accuracy. Evaluation is conducted in both white-box and black-box attack settings [1].

Datasets: We use two popular image classification datasets, MNIST [23], and CIFAR10 [20]. The pixel value of MNIST is a real number in [0, 1], and the pixel value of CFIAR10 is an integer in [0, 255].

Baseline attacks: For the two baseline attacks, CW is implemented using the source code [6] in [9] with the default configurations except that 1000 attack iterations are used to obtain stronger attack ability. PGD is implemented using source code [7] in [30] with the default settings. To enhance attack ability, we further combine SPA with the two baseline attacks and evaluate their effectiveness.

Baseline defenses: We evaluate SPA against two known defense techniques, PGD-based adversarial training [25] and randomized smoothing [13]. PGD-based adversarial training on CIFAR10 [7]and MNIST [8] is implemented with default parameters. PGD-based adversarial training has been demonstrated to be the most effective defense method against L_∞-norm based attacks on the MNIST and CIFAR10 datasets [3]. It can also be generalized to defend other L_p-norm attacks. Randomized smoothing is the most practically effective method to defend L_2-norm based attacks [2].

Target models: For MNIST, we use two variants of LeNet [23], namely *LeNet1* and *LeNet2*. LeNet1 is regarded as the primary network for MNIST in our experiments. For CIFAR10, we adopt WideResNet-32×10 [38] as the primary network and also use ResNet-32 [18] as a different network architecture for comparison. Both LeNet1 and WideResNet-32×10 have been used in the Madry's MNIST [8] and CIFAR10 Adversarial Examples Challenges [9], respectively. In each case, these models are trained both under naturally and adversarially setting.

To train a randomized smoothing model, we add a Gaussian noise layer with zero mean in front of the original target model and retrain it to get the randomized smoothing model. The standard deviation of the Gaussian noise is set to 1.0 and 64 (0.25×255) for MNIST and CIFAR10, respectively, following the settings in [13].

Parameter settings in SPA: For the MNIST dataset, we simply partition the pixel value range $[0, 1]$ into 255 equal-sized intervals, which is aligned with the standard image quantification and representation methods. For SPA based white-box attack, we set the maximum allowed perturbation size δ to 0.4, which is larger than the default value of 0.3 in PGD. The step size σ is set to 0.01 and the number of steps K is set to 40 (same as the default setting in PGD) in Algorithm 1. For SPA+PGD attack, we first generate a SPA adversarial image with

[5] Even though we consider SPA as a constrained optimization problem in the paper, SPA can also be formulated as an unconstrained problem with the L_p norm restriction included in the objective function as CW [9] and EAD [10] attacks.

[6] https://github.com/carlini/nn_robust_attacks
[7] https://github.com/ashafahi/free_adv_train
[8] https://github.com/MadryLab/mnist_challenge
[9] https://github.com/MadryLab/cifar10_challenge

$\delta = 0.4$ and then apply PGD attack with the default PGD setting except using a smaller perturbation size $\delta = 0.1$ to ensure that the generated image is natural. For SPA+CW attack, we follow the same SPA setting and the aforementioned CW setting.

For the CIFAR10 dataset, each image is partitioned into 3 sub-areas corresponding to the three channels (RGB). For each channel, the pixel value range is then evenly partitioned into 255 intervals, similar to the MNIST dataset. We set the maximum allowed perturbation size δ to 20, which is larger than the default value 8 used in PGD attack. We will show that despite the large difference in the perturbation size, the generated SPA images and PGD images are comparable in terms of how natural they are when viewed by human beings. In Algorithm 1, the step size σ is set to 2 and the number of steps K is set to 20 (both follow the same default setting in PGD attack). For SPA+PGD attack, we first generate an SPA adversarial image and then apply PGD attack with a smaller perturbation size $\delta = 2$. Again, we use the same SPA and CW setting for the SPA+CW attack.

It should be pointed out that one-shot attacks are inefficient for models protected by randomized transformations to the input as in the case of randomized smoothing models [3]. In this case, we adopt the approach of Expectation over Transformation (EOT) in [4] and compute the gradient over the expected transformation. We set the number of EOT to 50, following the setting in [2].

4.2 Evaluation results for white-box attacks

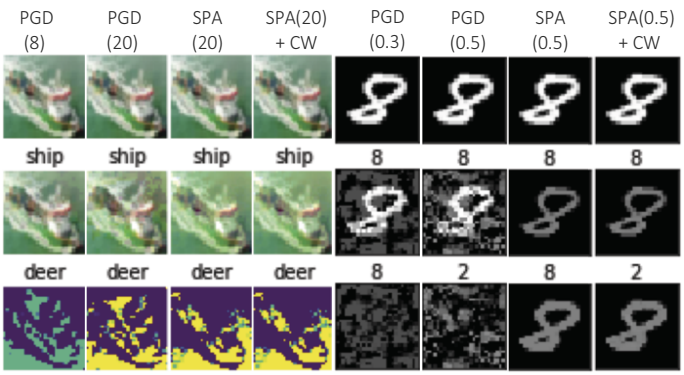

Figure 3: Examples of adversarial images generated by SPA, PGD and SPA+CW attacks. The three rows show the original images, the adversarial images and the corresponding perturbations, respectively, generated by the aforementioned attack models against the PGD-based adversarially trained primary networks on the CIFAR10 (left) and MNIST (right) dataset. The maximum allowed L_∞ perturbation sizes δ are shown in parentheses.

We report the evaluation results for white-box attacks in Figure 4. For PGD-based adversarial training, we show the results for both standard PGD distortion sizes (0.3 for MNIST and 8 for CIFAR10) and larger sizes (0.4 for MNIST and 20 also 22 for CIFAR10) for fair comparison. We note that when the LeNet1 model is PGD-adversarially trained with a distortion size beyond 0.5, the accuracy (on clean images) drops below 10%. We consider the reason is that the large unstructured perturbations have destroyed the structure of the images. Thus we do not report the results for more larger distortion size for PGD-based adversarial training.

It is observed that the performance of SPA+CW is the best against nearly all target models under white-box attack setting. CW is good at attacking deterministic networks including both naturally trained and adversarially trained networks. We note that CW is poor at attacking stochastic networks such as RSM. However, after integrating SPA, SPA+CW could ferociously attack the RSM.

We observe that SPA is superior to PGD when attacking PGD-based adversarially trained models as it allows a larger maximum perturbation size δ. Further, SPA is comparable to PGD for both naturally trained models and RSM. Moreover, by simply combining SPA with PGD, the revised model demonstrates much better attack ability than the original models. Thus, although the original SPA may not perform better than PGD and CW for all cases, it can be easily integrated with other attacks to obtain much better attack ability.

4.3 Black-box attacks and transferability

In this section, we present experimental results for black-box attacks (reported in Figure 5). In standard transfer-based black-box attacks (in contrast to query-based attacks [11, 6]), attackers first specify a substitute model to the black-box model, then generate a set of adversarial examples that could successfully attack the substitute model [26, 27, 24]. These generated adversarial examples are considered to have strong transferable attack ability and are consequently used to attack the target black-box model. For transfer-based black-box attacks, their effectiveness relies on how easily adversarial samples produced to mislead a specific model can also mislead other models [26]. Thus, the black-box attack ability in this case is determined by both white-box attack ability and transferability, where the latter depicts the accuracy consistency when images are tested with different models. To better analyze how the two factors influence black-box attack ability respectively, we disentangle them and define transferability as the reciprocal of the average absolute difference between the accuracy of a substitute model and that of a target model, where the average is taken over multiple target models for a fixed substitute model.

In order to demonstrate the black-box attack ability of SPA, we perform transfer-based black-box attacks across different target models evaluated on two datasets. In particular, the primary network for each dataset is used as the substitute model, and the rest networks are used as the target models (please refer to Section 4.1 for more details on network architectures).

From Figure 5, we observe that the transferability of SPA is generally higher than that of other attacks. In particular, the transferability of SPA on CIFAR10 is 28.50% higher than PGD, and 386.75% higher than CW. SPA consistently achieves low accuracy with or without defense and is extremely effective in the black-box setting. Although SPA does not perform significantly better than others against adversarially trained LeNet1 (with distortion size 0.4), the performance of SPA+PGD is satisfactory. Furthermore, SPA with a larger distortion size performs even better. Note that the black-box attack accuracy decreases to 54% when the distortion size increases to 0.5, while the distorted images can still be recognized very well as shown in Figure 3.

We notice that although PGD adversarial examples also exhibit high transferability, they are not as satisfactory as SPA in the black-box attack setting due to their relatively weak white-box attack performance on the substitute model.

Similar to white-box attacks, SPA+PGD exhibits better black-box attack ability than the original PGD and SPA. On the other hand, CW adversarial examples have demonstrated poor transferability and black-box attack ability in all the scenarios as shown in Figure 5. However, when combined with SPA, the black-box attack ability of SPA+CW improves significantly compared with CW itself. In particular, SPA+CW clearly outperforms all other attacks on CIFAR10.

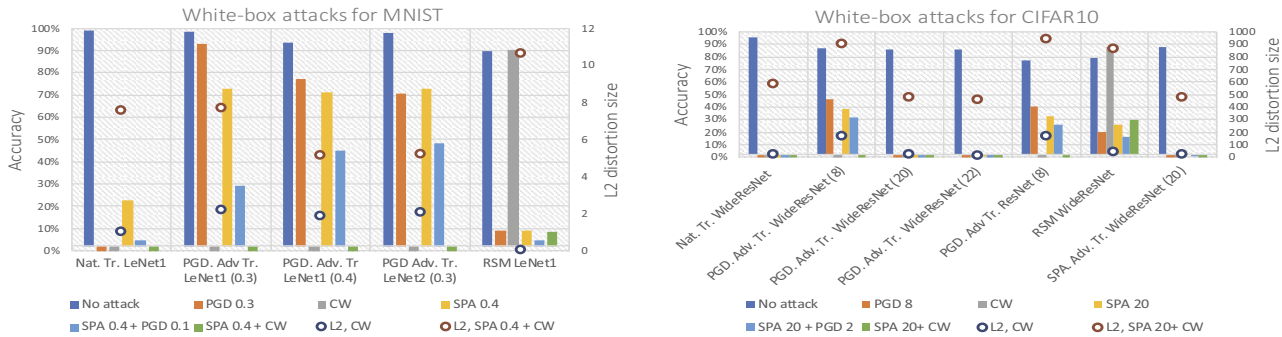

Figure 4: **White-box attacks**. Classification accuracy of the target networks under different attack-defense configurations on the MNIST (left) and CIFAR10 (right) datasets. "Nat. Tr." denotes the naturally trained network, "Adv. Tr" denotes the adversarially trained network and "RSM" represents the RSM model. The distortion sizes used in adversarial training are shown in parentheses.

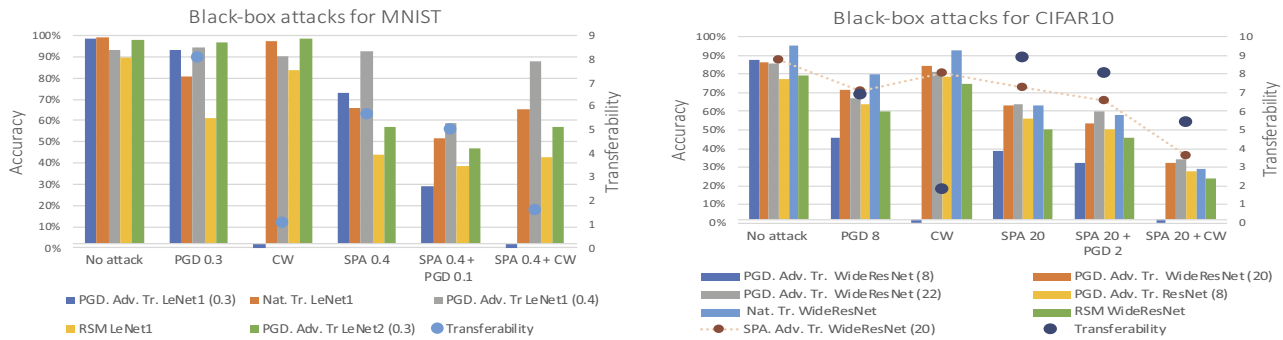

Figure 5: **Black-box attacks**. Classification accuracy on the MNIST (left) and CIFAR10 (right) datasets where the PGD-based adversarially trained primary networks LeNet1 and WideNetwork-32×10 are used as the substitute model for MNIST and CIFAR10, respectively. Other models are used as target models. The distortion sizes used in adversarial training are shown in parentheses.

4.4 The effect of attack space on attack ability

To simplify the discussion below, we first define two terms, namely, the *space flexibility* and the *distortion flexibility* of the attack space. Intuitively, the attack space denotes the flexibility that the attacker has in modifying images, which can be measured from two dimensions. The *space flexibility* refers to how many pixels in an image can be altered *independently* by the attacker, while the *distortion flexibility* measures to what extent each pixel can be modified (that is, the amount of perturbation that can be applied to each pixel). Traditional attacks independently twist each pixel, thus the space flexibility is $M \times N \times C$ (M, N and C are the height, width and the number of channels of an image, respectively). For our proposed structure based perturbation, all the pixels in a structure pattern are altered by the same amount, thus the number of pixels that can be changed independently (space flexibility) equals to the number of structure patterns $S \times V$, which is far smaller than that of the traditional attacks. On the other hand, our approach has greater distortion flexibility.

To understand how the *space flexibility* and *distortion flexibility* could affect the white- and black-box attack ability, we evaluate the performance of SPA by varying the distortion size ϵ and the interval size V (we fix S to 3 in this experiment). The results are plotted in Figures 6 and 7. For white-box attack (see Figure 6), the distortion size is superior to the interval size in affecting white-box attack ability, and larger distortion confers quite higher white-box attack ability. For black-box attack shown in Figure 7, it is observed that there exists a turning point of distortion size and interval size, and the transferability gradually converges after the turning point. However, considering the strong white-box attack ability brought by the large distortion size, the larger the distortion size is, the better the black-box attack ability can be achieved.

From these observations, we can infer that the distortion flexibility, rather than the space flexibility, plays a more important role in affecting white- and black-box attack ability. This is consistent with the *i.i.d.* assumption that traditional supervised learning models including DNNs all rely on, where models become much less effective under distribution-shift data compared to testing data follow the same distribution as training data. A large distortion size usually shifts data to a different data distribution, which invalidates target models on the distribution-shifted data. Therefore, it is desired to moderately sacrifice space flexibility to allow for more distortion flexibility with the purpose of achieving higher black- and white- box attack ability simultaneously. This is the core contribution of our SPA approach.

4.5 Illustration of adversarial examples

Figures 3 shows the adversarial examples generated by SPA, PGD and SPA+CW attacks (CW and SPA+PGD adversarial images are not shown due to space limitation). We observe that SPA adversarial examples indeed keep the structures of the original images. Although a larger maximum distortion is allowed than PGD, SPA adversarial examples are still clean and legible to humans. In particular, we observe that the SPA adversarial examples on MNIST are far more natural to human eyes than the PGD adversarial examples generated under a smaller perturbation size. Moreover, it is interesting to observe that the generated images remain natural even when SPA is combined with CW or PGD.

4.6 Adversarial training with SPA

Given the fact that adversarial training with PGD is less effective for SPA-based attacks, it is natural to wonder whether adversarial training with SPA is more effective. As SPA has far fewer parameters, it is reasonable to perform adversarial training with SPA. To this end, we

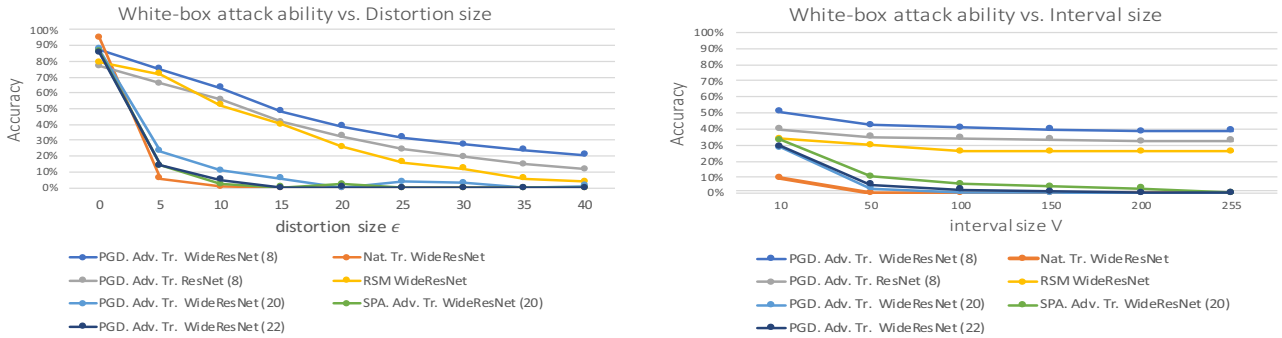

Figure 6: The effect of distortion size and interval size on white-box attack ability. With the increase of distortion size ϵ, the white-box attack ability consistently increases. However, enlarging interval size V only improves white-box attack ability marginally after a certain threshold.

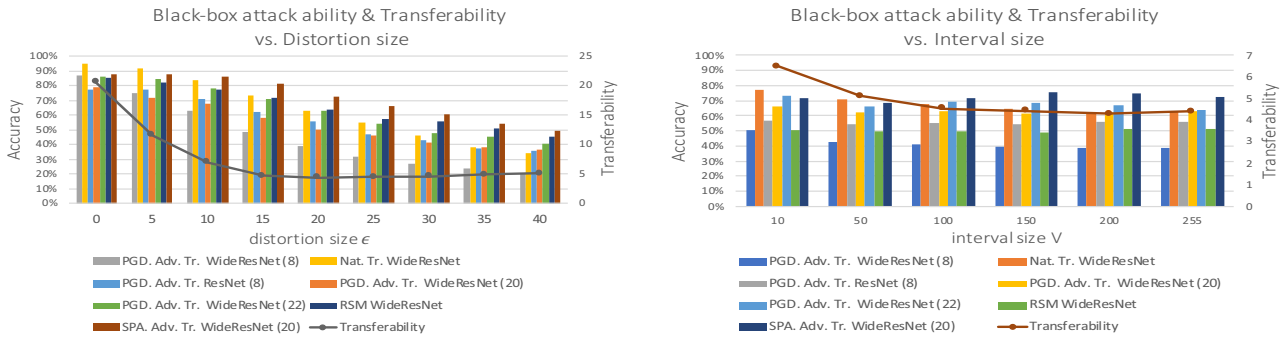

Figure 7: The effect of distortion size and interval size on black-box attack ability. The transferability saturated after a tuning point for both distortion and interval sizes. Similar to white-box attack, increasing distortion size ϵ profoundly raises the black attack ability, but there is little effect on black-box attack ability via increasing interval size.

conduct SPA-based adversarial training for the network WideResNet-32×10 on the CIFAR10 dataset and compare it with PGD-based adversarial training. We stress that it is extremely time-consuming to perform adversarial training by following the vanilla PGD-adversarial training paradigm. Inspired by the fast PGD-adversarial training proposed in [30], we design a fast SPA-adversarial training method by simultaneously computing the gradient with respect to the network weights and meta-perturbation η. We train WideResNet-32×10 using the following parameters: 80000 iterations, the batch-size is 128, the replay parameter is 4, and the perturbation size 20. From the results in Figures 4 and 5, we observe that SPA-based adversarial training does not achieve significant performance improvement for defending against SPA and other attacks even when compared with naturally trained models under white-box attacks. However, the SPA-based adversarially trained models have considerable black-box defense ability against SPA and the other two baseline attacks. Even so, SPA+CW can still satisfactorily attack the SPA-based adversarially trained model.

From the above results, we conclude SPA and its variants are profoundly effective in all the scenarios we have evaluated including both white-box and black-box settings.

5 RELATED WORK

There are a few recent works that focus on generating adversarial examples beyond the small L_p-norm perturbation restriction. In one direction, novel perturbation measures beyond L_p norms such as Wasserstein distance have been proposed [36]. In a different direction, small perturbations are imposed onto a latent representation of images instead of the images themselves [39, 32, 29]. Further, a growing line of work exploit domain knowledge to relax the small perturbation restriction while keeping the generated images natural and meaningful. Our work is aligned with this general framework. In particular, the work in [7] focuses on exploiting texture transfer and colorization to generate unrestricted images. In the context of physical-world attacks, perceivable perturbations that resemble real and inconspicuous objects have been proposed [15, 31].

Our SPA approach extends the Structure-Preserving Transformation (SPT) technique in our previous work [28], where we define a structure as a set of all pixels with the *same* pixel value, which is a special case of our definition when each interval contains a single pixel value. As SPT completely abandons the perturbation restriction, the uncontrolled excessive distortion may lead to unnatural adversarial examples. This drawback has been staved off in our SPA. There are a few recent works that also consider structural-aware perturbations similar to ours. In particular, color-shifted images are proposed in [19] where RGB images are first converted into the HSV color space and the hue and saturation components are then changed randomly where all the pixels in the same channel are perturbed by the same amount. This can be viewed as a special case of our approach where the pixel value partition has a single interval. In [37], structural perturbations are generated by penalizing the so called *group sparsity*. In contrast, our definition of structure patterns is better aligned with human intuition. Moreover, the attack in [37] can be viewed as a specific L_∞-bounded attack with an additional strong group sparsity restriction. Thus, it still suffers from the shortcomings of small-perturbation based attacks.

6 CONCLUSION

In this paper, we propose structure-preserving attack (SPA) as a new technique for generating natural and highly transferable adversarial examples. SPA is built upon an intuitive definition of structure patterns and introduces the concept of structural perturbation that relaxes the traditional small-perturbation requirement. Empirical results on the MNIST and CIFAR10 datasets show that SPA exhibits strong

attack ability in both the white-box and black-box settings even when defenses are applied. Further, when combined with PGD and CW attacks, SPA+PGD and SPA+CW exhibit even stronger white-box attack ability while retaining the good transferability of SPA.

We analyze the attack abilities of SPA and baseline attacks in terms of their space flexibility and distortion flexibility. The key insight is that it is beneficial to allow more distortion flexibility at the cost of space flexibility in order to achieve higher attack ability. We highlight that the high successful attack rates and the outstanding transferability of SPA stem from the fact that SPA exhibits greater distortion flexibility compared with traditional small-perturbation based approaches. By bridging the gap between the attacks that follow the strict small-perturbation restriction (extremely low distortion flexibility and extremely high space flexibility) and the attacks that allow unbounded distortions (extremely high distortion flexibility and low space flexibility), SPA opens up a new direction on generating natural and strong adversarial examples.

REFERENCES

[1] Naveed Akhtar and Ajmal Mian, 'Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey', *arXiv e-prints*, arXiv:1801.00553, (2018).

[2] Alexandre Araujo, Rafael Pinot, Benjamin Negrevergne, Laurent Meunier, Yann Chevaleyre, Florian Yger, and Jamal Atif, 'Robust neural networks using randomized adversarial training', *arXiv preprint arXiv:1903.10219*, (2019).

[3] Anish Athalye, Nicholas Carlini, and David Wagner, 'Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples', in *International Conference on Machine Learning*, (2018).

[4] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok, 'Synthesizing robust adversarial examples', *arXiv preprint arXiv:1707.07397*, (2017).

[5] S. Bahmani and B. Raj, 'A unifying analysis of projected gradient descent for ℓ_p-constrained least squares', *Applied and Computational Harmonic Analysis*, (2013).

[6] Arjun Nitin Bhagoji, Warren He, Bo Li, and Dawn Song, 'Practical black-box attacks on deep neural networks using efficient query mechanisms', in *European Conference on Computer Vision*, (2018).

[7] Anand Bhattad, Min Jin Chong, Kaizhao Liang, Bo Li, and David A Forsyth, 'Big but imperceptible adversarial perturbations via semantic manipulation', *arXiv preprint arXiv:1904.06347*, (2019).

[8] Nicholas Carlini and David Wagner, 'Adversarial examples are not easily detected: Bypassing ten detection methods', in *10th ACM Workshop on Artificial Intelligence and Security*, (2017).

[9] Nicholas Carlini and David Wagner, 'Towards evaluating the robustness of neural networks', in *IEEE Symposium on Security and Privacy*, (2017).

[10] Pin-Yu Chen, Yash Sharma, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh, 'Ead: elastic-net attacks to deep neural networks via adversarial examples', in *AAAI conference on artificial intelligence*, (2018).

[11] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh, 'Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models', in *10th ACM Workshop on Artificial Intelligence and Security*, (2017).

[12] Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier, 'Parseval networks: Improving robustness to adversarial examples', in *International Conference on Machine Learning*, (2017).

[13] Jeremy M Cohen, Elan Rosenfeld, and J Zico Kolter, 'Certified adversarial robustness via randomized smoothing', *arXiv preprint arXiv:1902.02918*, (2019).

[14] Gamaleldin F Elsayed, Shreya Shankar, Brian Cheung, Nicolas Papernot, Alex Kurakin, Ian Goodfellow, and Jascha Sohl-Dickstein, 'Adversarial examples that fool both human and computer vision', *arXiv preprint arXiv:1802.08195*, (2018).

[15] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song, 'Robust physical-world attacks on deep learning models', *arXiv preprint arXiv:1707.08945*, (2017).

[16] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy, 'Explaining and harnessing adversarial examples', in *International Conference on Learning Representations*, (2015).

[17] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten, 'Countering adversarial images using input transformations', in *International Conference on Learning Representations*, (2018).

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, 'Deep residual learning for image recognition', in *IEEE conference on computer vision and pattern recognition*, (2016).

[19] Hossein Hosseini and Radha Poovendran, 'Semantic adversarial examples', *arXiv preprint arXiv:1804.00499*, (2018).

[20] Alex Krizhevsky, 'Learning multiple layers of features from tiny images', *University of Toronto*, (2012).

[21] Alexey Kurakin, Ian Goodfellow, and Samy Bengio, 'Adversarial machine learning at scale', *arXiv preprint arXiv:1611.01236*, (2016).

[22] B. Landau, L. B. Smith, and S. S. Jones, 'The importance of shape in early lexical learning', *Cognitive Development*, (1988).

[23] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner, 'Gradient-based learning applied to document recognition', *IEEE*, (1998).

[24] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song, 'Delving into transferable adversarial examples and black-box attacks', *arXiv preprint arXiv:1611.02770*, (2016).

[25] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu, 'Towards deep learning models resistant to adversarial attacks', in *International Conference on Learning Representations*, (2018).

[26] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow, 'Transferability in machine learning: from phenomena to black-box attacks using adversarial samples', *arXiv preprint arXiv:1605.07277*, (2016).

[27] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami, 'Practical black-box attacks against machine learning', in *ACM on Asia conference on computer and communications security*, (2017).

[28] Dan Peng, Zizhan Zheng, and Xiaofeng Zhang, 'Structure-preserving transformation: Generating diverse and transferable adversarial examples', *arXiv preprint arXiv:1809.02786*, (2018).

[29] Haonan Qiu, Chaowei Xiao, Lei Yang, Xinchen Yan, Honglak Lee, and Bo Li, 'Semanticadv: Generating adversarial examples via attribute-conditional image editing', *arXiv preprint arXiv:1906.07927*, (2019).

[30] Ali Shafahi, Mahyar Najibi, Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein, 'Adversarial training for free!', *arXiv preprint arXiv:1904.12843*, (2019).

[31] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K Reiter, 'Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition', in *ACM SIGSAC Conference on Computer and Communications Security*, (2016).

[32] Yang Song, Rui Shu, Nate Kushman, and Stefano Ermon, 'Constructing unrestricted adversarial examples with generative models', in *Advances in Neural Information Processing Systems*, (2018).

[33] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai, 'Attacking convolutional neural network using differential evolution', *arXiv preprint arXiv:1804.07062*, (2018).

[34] Octavian Suciu, Scott E Coull, and Jeffrey Johns, 'Exploring adversarial examples in malware detection', in *IEEE Security and Privacy Workshops*, (2019).

[35] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus, 'Intriguing properties of neural networks', in *International Conference on Learning Representations*, (2014).

[36] Eric Wong, Frank R Schmidt, and J Zico Kolter, 'Wasserstein adversarial examples via projected sinkhorn iterations', *arXiv preprint arXiv:1902.07906*, (2019).

[37] Kaidi Xu, Sijia Liu, Pu Zhao, Pin-Yu Chen, Huan Zhang, Quanfu Fan, Deniz Erdogmus, Yanzhi Wang, and Xue Lin, 'Structured adversarial attack: Towards general implementation and better interpretability', in *International Conference on Learning Representations*, (2019).

[38] Sergey Zagoruyko and Nikos Komodakis, 'Wide residual networks', *arXiv preprint arXiv:1605.07146*, (2016).

[39] Zhengli Zhao, Dheeru Dua, and Sameer Singh, 'Generating natural adversarial examples', in *International Conference on Learning Representations*, (2018).