BREAKING CERTIFIED DEFENSES: SEMANTIC ADVERSARIAL EXAMPLES WITH SPOOFED ROBUSTNESS CERTIFICATES

Anonymous authors
Paper under double-blind review

ABSTRACT

Defenses against adversarial attacks can be classified into certified and non-certified. Certifiable defenses make networks robust within a certain ℓ_p-bounded radius, so that it is impossible for the adversary to make adversarial examples in the certificate bound. We present an attack that maintains the imperceptibility property of adversarial examples while being outside of the certified radius. The proposed "Shadow Attack" can fool certifiably robust networks while simultaneously producing a strong "spoofed" certificate.

1 INTRODUCTION

Conventional training of natural networks has been shown to produce classifiers that are highly sensitive to imperceptible adversarial perturbations (Szegedy et al., 2013; Biggio et al., 2013), "natural looking" images that have been manipulated to causing misclassified by a neural network (Figure 1). While a wide range of defenses exist that harden neural networks against such attacks (Madry et al., 2017; Shafahi et al., 2019), many attacks based on heuristics and tricks have been shown to be easily breakable Athalye et al. (2018). This has motivated work on *certifiably* secure networks — classifiers that produce a classification, and also (when possible) a rigorous guarantee that the input is not adversarially manipulated (Cohen et al., 2019; Zhang et al., 2019b).

To date, all work on certifiable defenses has focused on deflecting ℓ_p-bounded attacks, where $p = 2$ or ∞ (Cohen et al., 2019; Gowal et al., 2018; Wong et al., 2018). After labelling an images, these defenses then check that there exists an image with a different label within an ℓ_p ball of radius ϵ centered on the input image, where ϵ is a security parameter chosen by the user. If no such image exists, then they certify that the input image is not a ℓ_p adversarial example.

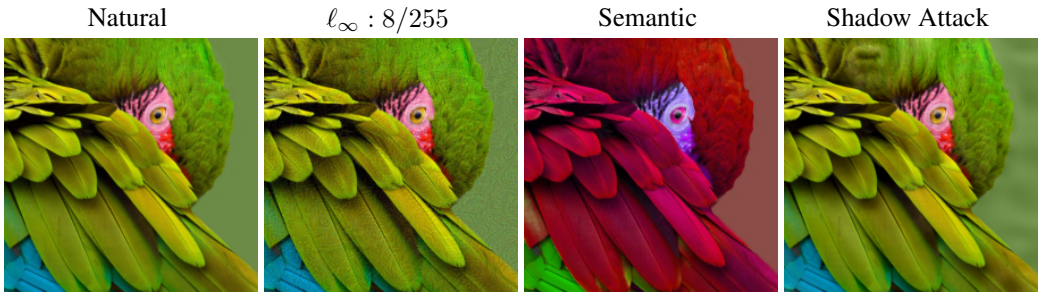


Figure 1: All adversarial examples have the goal of fooling classifiers while looking "natural". The ℓ_p-bounded attacks limit the adversarial perturbation pixel values while semantic attacks are unrestricted in terms of ℓ_p-norms but they also produce natural looking images. Our Shadow Attack falls within the category of unrestricted attacks and is assigned a large certified radii by the smoothed certified classifier it is attacking. Our attack is not ℓ_p-bounded and the example has $\ell_\infty(\delta) = 17.6/255$ and $\ell_2(\delta) = 5.5$.

In this work, we demonstrate how a system that relies on certificates as a measure of label security can be exploited. We present a new class of adversarial examples that target not only the classifier output label, *but also the certificate.* We do this by adding adversarial perturbations to images that are large in the ℓ_p norm (larger than the ϵ used by the certificate generator), and result in an attack image that is surrounded by a large ℓ_p ball exclusively containing images of the same (adversarially chosen) label. The resulting attacks produce a "spoofed" certificate with a seemingly strong security guarantee despite being adversarially manipulated. Note that the statement made by the certificate (i.e., that the input image is not an ϵ adversarial example in chosen norm) is still technically correct, however in this case the adversary is hiding behind a certificate to avoid detection by a certifiable defense.

In summary, we consider methods that attack a certified classifier in the following sense:

- **Imperceptibility:** the adversarial example is "natural-looking" or "looks like" its corresponding natural example,
- **Misclassification:** the certified classifier assigns an incorrect label to the adversarial example, and
- **Strongly certified:** the certified classifier provides a strong/large-radius certificate for the adversarial example.

While the existence of such an attack does not invalidate the certificates produced by certifiable systems, it should serve as a warning that certifiable defenses are not inherently secure, and one should not strongly rely on them as an indicator of label correctness.

BACKGROUND

In the white-box setting, where the attacker knows the victim's network and parameters, adversarial perturbation can often be constructed using first order gradient information (Carlini & Wagner, 2017; Kurakin et al., 2016; Moosavi-Dezfooli et al., 2016) or using approximations of the gradient (Uesato et al., 2018; Athalye et al., 2018). The prevailing optimization formulation for crafting adversarial examples uses an additive adversarial perturbation, and perceptibility is minimized using an ℓ_p-norm constraint. Different ℓ_p-norms result in different measures of imperceptibility. For example, ℓ_∞-bounded adversarial attacks limit every pixel from being large , while ℓ_0 adversarial attacks do not limit every individual pixel but limit the number of pixels that can be modified (Wiyatno & Xu, 2018).

Limiting an adversarial image to have a small ℓ_p-norm is not the only way of maintaining natural looking attacks. Consequently, unrestricted and non-bounded adversarial examples have been subject of recent studies (Brown et al., 2018). For example, Hosseini & Poovendran (2018) use shifting color channels, and Engstrom et al. (2017) use rotation and translation to craft "semantic" adversarial examples. In Figure 1, we produce the semantic adversarial examples using the method of Hosseini & Poovendran (2018) which is a greedy approach that transforms the image into HSV space, and then, while keeping V constant, tries to find the smallest S perturbation causing misclassification[1]. Other variants use generative models to construct natural looking images causing misclassification (Song et al., 2018; Dunn et al., 2019).

In practice, many of the defenses which top adversarial defense leader-board challenges are non-certified defenses (Madry et al., 2017; Zhang et al., 2019a; Shafahi et al., 2019). The majority of these defenses make use of adversarial training which is the process of training on adversarial examples built for the network being trained. These non-certified defenses are mostly evaluated against PGD-based attacks, which results in an upper-bound on robustness.

Certified defenses, on the other-hand, provably make networks resist ℓ_p-bounded perturbations of a certain radius. For instance, randomized smoothing (Cohen et al., 2019) is a certifiable defense against ℓ_2-norm bounded attacks, and CROWN-IBP (Zhang et al., 2019b) is a certifiable defense against l_∞-norm bounded perturbations.

[1] In fig. 1, the adversarial example has saturation=0

To the best of our knowledge, prior works have focused on making adversarial examples that satisfy the imperceptibility and misclassification conditions, but none have investigated manipulating certificates, which is our focus here.

The reminder of this paper is organized as follows. In section 2 we introduce our new approach Shadow Attack for generating adversarial perturbations. In section 3 we show our results on attacking "randomized Smoothing" certificates (Cohen et al., 2019). Using "randomized smoothing" as an example, in section 4 we do an ablation study and show why the elements of the Shadow Attack are important for successfully manipulating certified models. In section 5 we generate adversarial examples for "CROWN-IBP" (Zhang et al., 2019b). Finally, we discuss results and wrap up in section 6.

2 THE SHADOW ATTACK

Because certificate spoofing requires large perturbations (larger than the ℓ_p ball of the certificate), we propose a simple attack that enables numerous modes for creating large perturbations. Our attack can be seen as the generalization of the well-known PGD attack, which creates adversarial images by perturbing a clean base image. Given a loss function L and an ℓ_p-norm bound ϵ for some $p \geq 0$, PGD attacks solve the following optimization problem:

$$\max_{\delta} L(\theta, x + \delta) \tag{1}$$

$$\textit{s.t. } \|\delta\|_p \leq \epsilon, \tag{2}$$

where θ are the network parameters and δ is the adversarial perturbation to be added to the clean input image x. Constraint 2 promoted imperceptibility of the resulting perturbation to the human eye by limiting the perturbation size. In the shadow attack, instead of solving the above optimization problem, we solve the following:

$$\max_{\delta} L(\theta, x + \delta) - \lambda_c C(\delta) - \lambda_{tv} TV(\delta) - \lambda_s Sim(\delta), \tag{3}$$

where $\lambda_c, \lambda_{tv}, \lambda_s$ are scalar penalty weights.

Constraint $TV(\delta)$ forces the perturbation δ to have a small total variation (TV), and so be smooth. This constraint forces the perturbation to appear more like a natural image, given that natural images often have small TV. Constraint $C(\delta)$ limits the perturbation δ globally by constraining the change in the mean of each color channel c. This constraint is needed since total variation is invariant to constant/scalar additions to each color channel, and it is desirable to supress extreme changes in the color balances of images.

Constraint $Sim(\delta)$ enforces the perturbation δ to have similar values in each color channel.In the case of an RGB image of shape $3 \times W \times H$, if $Sim(\delta)$ is small, the perturbation roughly adds the same amount to the Red, Green, and Blue channels at every spatial pixel location: $\delta_{R,w,h} \approx \delta_{G,w,h} \approx \delta_{B,w,h}, \quad \forall (w,h) \in W \times H$. Adding/subtracting the same amount to RGB channels results in making the corresponding pixels darker/lighter, without changing the color balance of the image.

Later, in section 3, we suggest two ways of enforcing such similarity between RGB channels and we find both of them effective:

- **1-channel attack:** strictly enforces $\delta_{R,i} \approx \delta_{G,i} \approx \delta_{B,i}, \quad \forall i$ by using a single-channel matrix to represent $\delta_{W \times H}$ and duplicate δ to make a 3-channel image. In this case, $Sim(\delta) = 0$, and perturbation is greyscale.
- **3-channel attack:** use a 3-channel perturbation $\delta_{3 \times W \times H}$ and define a function that measures dissimilarity such as: $Sim(\delta) = \|\delta_R - \delta_B\|_p + \|\delta_R - \delta_G\|_p + \|\delta_B - \delta_G\|_p$.

All together, the three constraints enforce the perturbation to be (a) small (b) smooth and (c) without dramatic color changes (e.g. swapping blue to red).

2.1 SPOOFING A CERTIFICATE

The goal of our attack is not only to find natural looking images that misclassify, but also to generate strong certificates. To achieve this goal, we need a loss penalty that promotes an increased certification radius.

We focus on spoofing certificates for *untargeted* attacks in which the attacker does not specify the class into which the attack image moves. In the untargeted case, we can generate an adversarial perturbation for all possible wrong classes $\bar{y}$ and choose the best one as our strong attack:

$$\max_{\bar{y} \neq y, \delta} -L(\theta, x + \delta \| \bar{y}) - \lambda_c C(\delta) - \lambda_{tv} TV(\delta) - \lambda_s Sim(\delta) \tag{4}$$

where y is the true label/class for the clean image x, and L is a spoofing loss that promotes a large certificate. We examine different choices for L for different certificates below.

3 ATTACKS ON RANDOMIZED SMOOTHING

The randomized Smoothing method, first proposed by Lecuyer et al. (2018) and later improved by Li et al. (2018), is an adversarial defense against l_2-norm bounded attacks. Cohen et al. (2019) prove a tight robustness gaurantee under the ℓ_2 norm for smoothing with gaussian noise. Their study was the first certifiable defense for the ImageNet dataset (Deng et al., 2009). The method constructs certificates by first creating many copies of an input image contaminated with random Gaussian noise of standard deviation σ. Then, it uses a base classifier (a neural net) to make a prediction for all of the images in the randomly augmented batch. Depending on the level of the consensus of the class labels at these random images, a certified radius is calculated that can be at most 4σ.

Intuitively, if the image is far away from the decision boundary, the base classifier should predict the same label for each noisy copy of the test image, in which case the certificate is strong. On the other hand, if the image is adjacent to the decision boundary (i.e. it is easy for the adversary to generate adversarial perturbations), the base classifier's predictions for the Guassian augmented copies may vary ($x \hat{+} g_i \neq x \hat{+} g_j, \quad g \sim \mathcal{N}(0, \sigma^2)$ for augmented copies i & j). If the variation is large, the smoothed classifier abstains from making a prediction.

Consequently, for an adversary to engineer strong certificates for wrong classes (produce a large certified radius), they must make sure that the majority of a batch of Gaussian augmented images around the adversarial image vote for the same wrong label. Predicting the same label motivates the use of targeted attacks. Therefore, we try to minimize the cross-entropy given a wrong label instead of maximizing the cross-entropy loss which is common for non-targeted attacks.

One could modify the optimization problem of equation 4 to accommodate the randomized smoothing as:

$$\max_{\bar{y} \neq y, \delta} -L_{batch}(\theta, x + \delta \| \bar{y}) - \lambda_c C(\delta) - \lambda_{tv} TV(\delta) - \lambda_s Sim(\delta) \tag{5}$$

We note that L_{batch} refers to the average cross-entropy over the randomized Gaussian augmented batch of copies of x (b), and θ are the parameters of the base classifier which is trained with Gaussian data augmentation:

$$L_{batch} = \frac{1}{|b|} \sum_{x_i \in b} L(\theta, x_i + \delta \| \bar{y}) \tag{6}$$

The adversaries optimization problem aims to find a *targeted* universal adversarial perturbation for all the Gaussian augmented copies of the clean image x in batch b (Moosavi-Dezfooli et al., 2017). We follow Shafahi et al. (2018) to generate the universal perturbation.

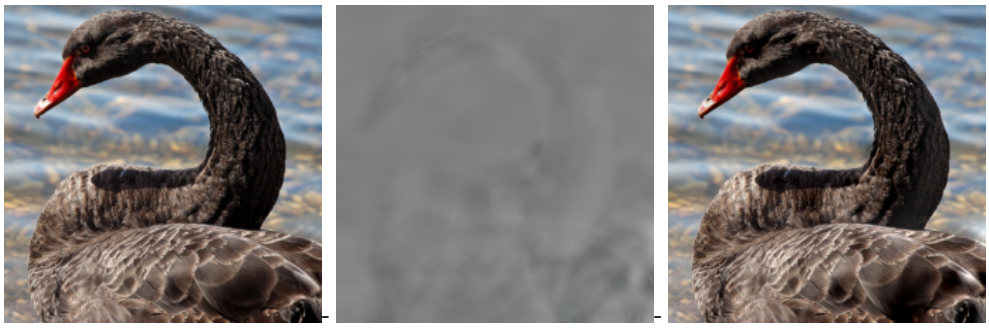
RESULTS

Cohen et al. (2019) show the performance of the Gaussian smoothed classifier on CIFAR-10 (Krizhevsky et al.) and ImageNet (Deng et al., 2009). To attack the CIFAR and ImageNet smoothed classifiers, we use $|b| = 400$, $\lambda_{tv} = 0.3$, $\lambda_c = 1.0$, and perform 300 steps of SGD with *step-size* $= 0.1$.

Our choices for the functional regularizers in constraints $C(\delta)$ and $TV(\delta)$ are:

Table 1: Shadow Attack results on Randomized Smoothing.

Dataset	$\sigma(l_2)$	**Randomized Smoothed**		**Shadow Attack**	
		Mean	STD	Mean	STD
CIFAR-10	0.12	0.14	0.056	**0.22**	0.005
	0.25	0.30	0.111	**0.35**	0.062
	0.50	0.47	0.234	**0.65**	0.14
	1.00	0.78	0.556	**0.85**	0.442
ImageNet	0.25	0.30	0.109	**0.31**	0.109
	0.50	0.61	0.217	0.38	0.191
	1.00	1.04	0.519	0.64	0.322



(a) Natural image (x) (b) Adversarial perturbation (δ) (c) Adversarial example ($x + \delta$)

Figure 2: The adversarial example built using our Shadow Attack for the smoothed ImageNet classifier for which the certifiable classifier produces a large certified radii and its corresponding adversarial perturbation. The adversarial noise is smooth and natural looking even-though it is large when measured using ℓ_p-metrics. True class: black swan; misclassified as: hook. Also see appendix 16.

$$C(\delta) = \|\text{Avg}(|\delta_R|), \text{Avg}(|\delta_G|), \text{Avg}(|\delta_B|)\|_2^2 \quad , \qquad TV(\delta_{i,j}) = \text{anisotropic-TV}(\delta_{i,j})^2$$

where $|\cdot|$ is the element-wise absolute value operator, and Avg computes the average. For the Sim regularizer, we experiment with both the (a) 1-Channel attack that ensures always $Sim(\delta) = 0$ and the (b) 3-Channel attack by setting $Sim(\delta) = \|(\delta_R - \delta_G)^2, (\delta_R - \delta_B)^2, (\delta_G - \delta_B)^2\|_2$ and $\lambda_s = 0.5$. For the validation examples which the smooth classifier does not abstain (see Cohen et al. (2019) for more details), the less-constrained 3-channel attack is always able to find an adversarial example while the 1-channel attack performs great as well and achieves 98.5% success. In section 4 we will discuss in more details other differences between 1-channel and 3-channel attacks. The results are summarized in Table 1. For the various base-models and choices of σ, our adversarial examples are able to produce certified radii which are on average larger than the certified radii produced for their natural parallel. For ImageNet, since attacking all 999 remaining target classes were computationally intractable, we only attacked target class ids 100, 200, 300, 400, 500, 600, 700, 800, 900, and 1000.

Figure 2 depicts a sample adversarial example built for the smoothed ImageNet classifier which produces a large certificate. The adversarial noise is a universal noise which causes the batch of gaussian augmented black swan images to get misclassified as hooks. For more, see appendix 16.

4 Ablation Study of the Attack Parameters

In this section we perform ablation study on the parameters of Shadow Attack to evaluate (a) required PGD steps to find a successful attack, (b) the importance of λ_{Sim} or alternatively using 1-channel attacks, and last but not least, (c) the effect of λ_{tv}.

The default parameters for all of the experiments are as follows unless explicitly mentioned: We use $n = 30$ SGD steps and with step-size $s = 0.1$ for the optimization. All experiments except part (b) use 1-channel attacks for the sake of simplicity and efficiency (since it has less parameters). We

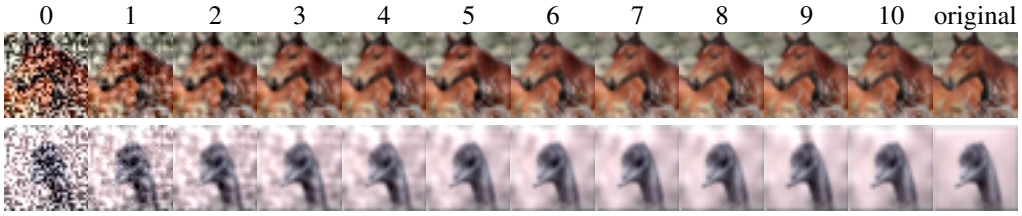


Figure 3: The first 10 steps of the optimization vs the original image.

assume $\lambda_{tv} = 0.3$, $\lambda_c = 20.0$ and universal batch-size $|b| = 50$. The dataset we use is a subset of CIFAR-10 dataset which including one example per each class (we selected the first example from each class in the CIFAR-10 evaluation data).

Figure 3 shows how the adversarial example evolves during the first few steps of the optimization (See appendix 13 for more examples). Also, figures 4, 5, and 6, show the average $L_b(\delta)$, $TV(\delta)$, and $C(\delta)$ respectively (Note that we use 1-channel attacks, so $Sim(\delta)$ is always 0). We empirically show that taking a few (even $n = 10$) pgd-steps is enough for the sake of convergence for CIFAR-10, but for our main results (i.e. attacking Randomized Smoothing in section 3 and attacking CROWN-IBP in section 5) we take $n = 300$ steps to be safe.

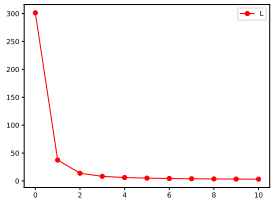


Figure 4: Average $L_b(\delta)$ in the first 10 steps.

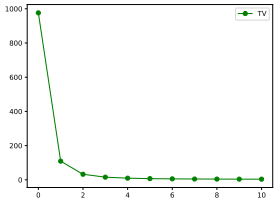


Figure 5: Average $TV(\delta)$ in the first 10 steps.

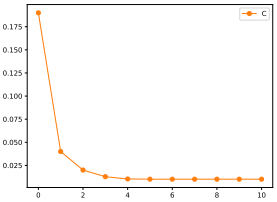


Figure 6: Average $C(\delta)$ in the first 10 steps.

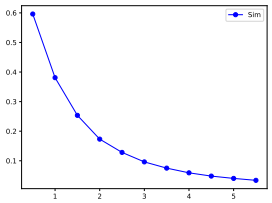


Figure 7: The effect of λ_{Sim} on the resulting $Sim(\delta)$

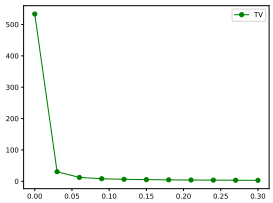


Figure 8: The effect of λ_{tv} on the resulting $TV(\delta)$

To explore the importance of λ_{sim}, we use 3-channel attacks and vary λ_{sim} to produce different images in figure 11[2].

Also, figure 7 shows the mean $Sim(\delta)$ for different values of $\lambda_{sim} (0 \leq \lambda_{sim} \leq 5.0)$. We also plot the histogram of the certificate radii in figure 9. Figure 10 compares 1-Channel vs 3-Channel attacks resulting images for some of randomly selected CIFAR-10 images.
Last but not least, we explore the effect of λ_{tv} on imperceptibility of the perturbations in Figure 12. See table 15 for more images and 8 to see it's effects on $TV(\delta)$.

[2]See appendix 14 for more images.

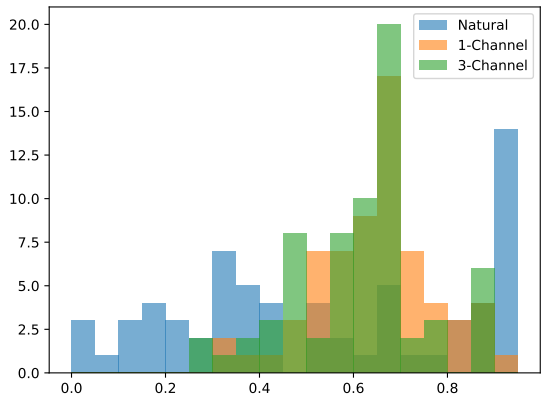


Figure 9: Histogram of randomized smoothed certificate radii for 100 randomly sampled CIFAR-10 validation images vs those calculated for their adversarial examples crafted using our 1-channel and 3-channel adversarial Shadow Attack attacks. The attacked base classifier used for the smoothed classifier is the Resnet-110 with $\sigma = 0.50$. 1-channel attacks are almost as good as the less-restricted 3-channel attacks.

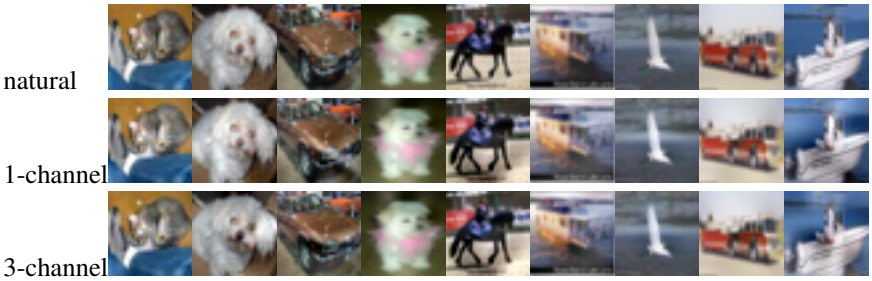


Figure 10: The visual effect of Shadow Attack on 9 randomly selected CIFAR-10 examples using 1-Channel and 3-Channel attacks.

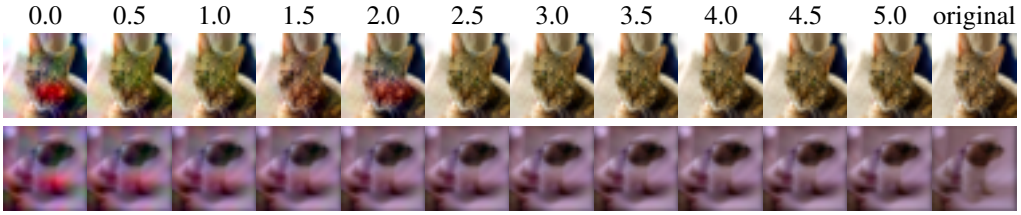


Figure 11: The visual effect of λ_{Sim} on imperceptibility of the perturbations. The first row indicates the value of λ_{Sim}.

Figure 12: The visual effect of λ_{tv} on the on imperceptibility of the perturbations. The first row shows the value of λ_{tv}

Table 2: Errors calculated based on the estimated distance to the boundary on natural and adversarial examples produces by Shadow Attack for the CIFAR-10 CROWN-IBP released models. Smaller is better.

$\epsilon(l_\infty)$	**Model Family**	**Method**	**Robustness Errors**		
			Min	Mean	Max
$2/255$	9 small models	CROWN-IPB	52.46	57.55	60.67
		Shadow Attack	68.91	74.77	82.09
	8 large models	CROWN-IBP	52.52	53.9	56.05
		Shadow Attack	79.7	82.28	85.09
$8/255$	9 small models	CROWN-IBP	71.28	72.15	73.66
		Shadow Attack	**66.14**	**69.3**	**72.12**
	8 large models	CROWN-IBP	70.79	71.17	72.29
		Shadow Attack	**67.9**	**70.36**	73.03

5 ATTACKS ON CROWN-IBP

Interval Bound Propagation (IBP) methods have been recently studied as a defense against ℓ_∞-bounded attacks. Many recent studies such as Gowal et al. (2018); Xiao et al. (2018); Wong et al. (2018); Mirman et al. (2018) have investigated IBP methods to train provably robust networks. To the best of our knowledge, the CROWN-IBP method by Zhang et al. (2019b) achieves state-of-the-art performance for MNIST (LeCun & Cortes, 2010), Fashion-MNIST (Xiao et al., 2017), and CIFAR-10 datasets among certifiable ℓ_∞ defenses. In this section we focus on attacking Zhang et al. (2019b) using CIFAR-10.

IBP methods estimate how much a small ℓ_∞-bounded noise in the input can propagate into the classification layer. By estimating the propagated error, they can find provable bounds for robustness in the input layer. To train robust networks, IBP methods include a term in their loss function that encourages the estimated bound on the training data from the decision boundary to increase.

Given that our attack also requires the estimated bound to be large, we directly use the loss function used in IBP training methods to attack the CROWN-IBP pre-trained models from Zhang et al. (2019b).

We attack 4 classes of networks released by the CROWN-IBP creators for CIFAR-10. There are two classes for IBP architectures, one of them consists of 9 small models and the other consists of 8 larger models. For each class of architectures, there are two sets of pre-trained models: one for $\epsilon = 2/255$ and one for $\epsilon = 8/255$. We use the same hyper-parameters and regularizers as in 3. For the sake of efficiency, we only do 1-channel attacks. We attack the 4 classes of models and for each class, we report the min, mean, and max of the robustness errors and compare them with those of the CROWN-IBP paper.

In the IBP literature, an "error" refers to an example that either has been misclassified or has been correctly classified but with an estimated robustness bound less than ϵ. We use a similar definition for robustness error, so that if misclassification does not happen or if the estimated distance to the boundary is smaller than ϵ, we count it as an error for the attack. Table 2 shows the results for each set of experiments. For the CROWN-IBP models trained on $\epsilon = 8/255$, our attack is capable of finding adversarial examples resulting in stronger certificates (i.e. smaller robustness errors) that natural images.

6 CONCLUSION

We demonstrate that it is possible to produce adversarial examples with strong certified robustness by using large-norm perturbations. This work suggests that the certificates produced by certifiably robust classifiers, which mathematically rigorous, are not always good indicators of robustness or accuracy. Our adversarial examples are built using our Shadow Attack which produces smooth and natural looking perturbations that are often less perceptible than those of the commonly used loosely-norm-bounded ℓ_p perturbations, while being large enough in norm to escape the certification regions of state of the art principled defenses.

REFERENCES

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420*, 2018.

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 387–402. Springer, 2013.

Tom B Brown, Nicholas Carlini, Chiyuan Zhang, Catherine Olsson, Paul Christiano, and Ian Goodfellow. Unrestricted adversarial examples. *arXiv preprint arXiv:1809.08352*, 2018.

Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pp. 3–14. ACM, 2017.

Jeremy M Cohen, Elan Rosenfeld, and J Zico Kolter. Certified adversarial robustness via randomized smoothing. *arXiv preprint arXiv:1902.02918*, 2019.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

Isaac Dunn, Tom Melham, and Daniel Kroening. Generating realistic unrestricted adversarial inputs using dual-objective gan training. *arXiv preprint arXiv:1905.02463*, 2019.

Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. A rotation and a translation suffice: Fooling cnns with simple transformations. *arXiv preprint arXiv:1712.02779*, 2017.

Sven Gowal, Krishnamurthy Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Timothy Mann, and Pushmeet Kohli. On the effectiveness of interval bound propagation for training verifiably robust models. *arXiv preprint arXiv:1810.12715*, 2018.

Hossein Hosseini and Radha Poovendran. Semantic adversarial examples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1614–1619, 2018.

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). URL `http://www.cs.toronto.edu/~kriz/cifar.html`.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

Cassidy Laidlaw and Soheil Feizi. Functional adversarial attacks. *arXiv preprint arXiv:1906.00001*, 2019.

Yann LeCun and Corinna Cortes. MNIST handwritten digit database. 2010. URL `http://yann.lecun.com/exdb/mnist/`.

Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. *arXiv preprint arXiv:1802.03471*, 2018.

Bai Li, Changyou Chen, Wenlin Wang, and Lawrence Carin. Second-order adversarial attack and certifiable robustness. *CoRR*, abs/1809.03113, 2018. URL `http://arxiv.org/abs/1809.03113`.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Matthew Mirman, Timon Gehr, and Martin Vechev. Differentiable abstract interpretation for provably robust neural networks. In *International Conference on Machine Learning*, pp. 3575–3583, 2018.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. 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.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1765–1773, 2017.

Ali Shafahi, Mahyar Najibi, Zheng Xu, John Dickerson, Larry S Davis, and Tom Goldstein. Universal adversarial training. *arXiv preprint arXiv:1811.11304*, 2018.

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.

Yang Song, Rui Shu, Nate Kushman, and Stefano Ermon. Constructing unrestricted adversarial examples with generative models. In *Advances in Neural Information Processing Systems*, pp. 8312–8323, 2018.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Jonathan Uesato, Brendan O'Donoghue, Aaron van den Oord, and Pushmeet Kohli. Adversarial risk and the dangers of evaluating against weak attacks. *arXiv preprint arXiv:1802.05666*, 2018.

Rey Wiyatno and Anqi Xu. Maximal jacobian-based saliency map attack. *arXiv preprint arXiv:1808.07945*, 2018.

Eric Wong, Frank Schmidt, Jan Hendrik Metzen, and J Zico Kolter. Scaling provable adversarial defenses. In *Advances in Neural Information Processing Systems*, pp. 8400–8409, 2018.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Kai Y Xiao, Vincent Tjeng, Nur Muhammad Shafiullah, and Aleksander Madry. Training for faster adversarial robustness verification via inducing relu stability. *arXiv preprint arXiv:1809.03008*, 2018.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P Xing, Laurent El Ghaoui, and Michael I Jordan. Theoretically principled trade-off between robustness and accuracy. *arXiv preprint arXiv:1901.08573*, 2019a.

Huan Zhang, Hongge Chen, Chaowei Xiao, Bo Li, Duane Boning, and Cho-Jui Hsieh. Towards stable and efficient training of verifiably robust neural networks. *arXiv preprint arXiv:1906.06316*, 2019b.

A APPENDIX

In this section, we include the complete results of our ablation study. As we mentioned in section 4, we use is a subset of CIFAR-10 dataset, including one example per each class. For the sake of simplicity, we call the dataset Tiny-CIFAR-10. Here, we show the complete results for the ablation experiments on all of Tiny-CIFAR-10 examples. Figure 13 shows that taking a few optimization steps is enough for the resulting images to look natural-looking. Figure 14 and 15, respectively show the effect of λ_{Sim} and λ_{TV} on the imperceptability of the perturbations.

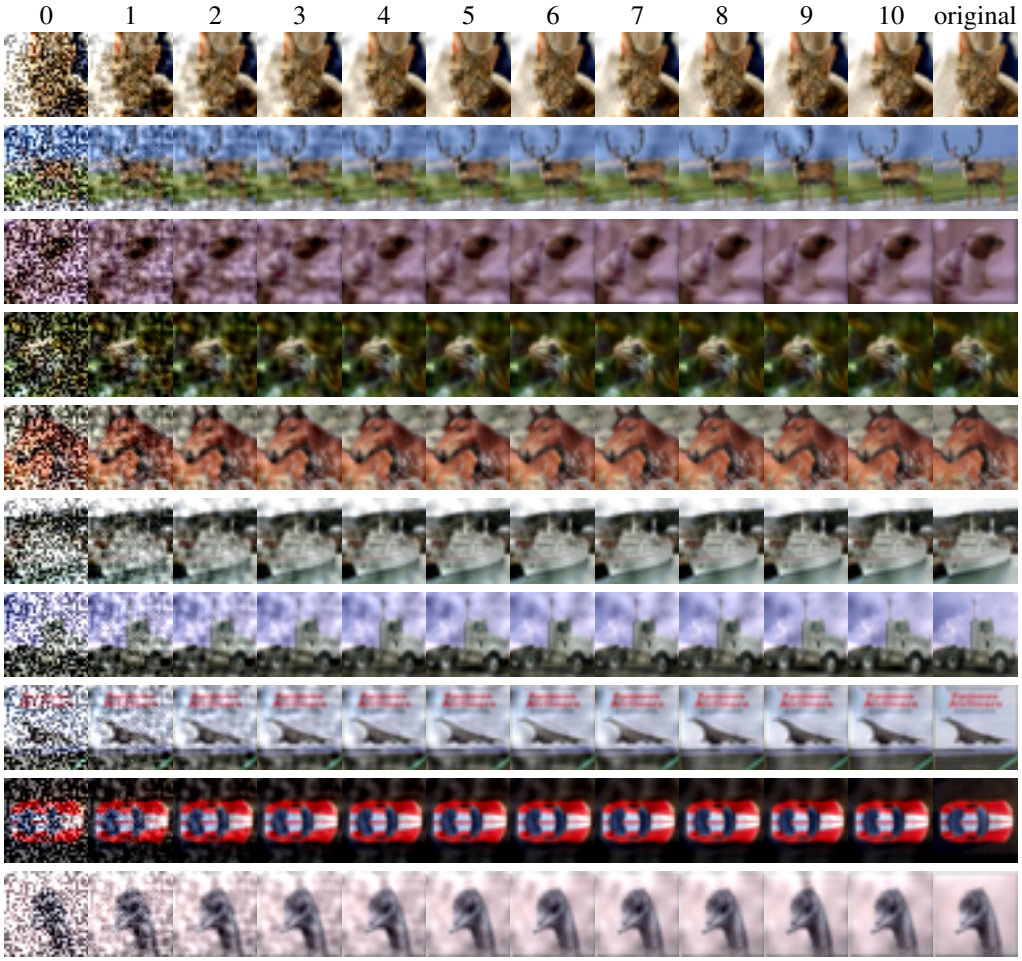


Figure 13: The first 10 steps of the optimization vs the original image for Tiny-CIFAR-10. See section 4 for the details of the experiments.

Figure 14: The visual effect of λ_{Sim} on $Sim(\delta)$ on Tiny-CIFAR-10. See section 4 for the details of the experiments.

Figure 15: The visual effect of λ_{tv} on the perturbation Tiny-CIFAR-10. See section 4 for the details of the experiments.

IMAGENET RESULTS

Many of the recent studies have explored the semantic attacks. Semantic attacks are powerful for attacking defenses (Engstrom et al., 2017; Hosseini & Poovendran, 2018; Laidlaw & Feizi, 2019). Many of semantic attacks are applicable to Imagenet, however, none of them consider increasing the radii of the certificates generated by the certifiable defenses.

Some other works focus on using generative models to generate adversarial examples (Song et al., 2018), but unfortunately none of the GAN's are expressive enough to capture the manifold of the ImageNet.

In this section, we show some of the successful examples generated by Shadow Attack to attack Randomized Smoothed classifiers for Imagenet.

Figure 16: Natural looking Imperceptible ImageNet adversarial images which produce large certified radii for the ImageNet Gaussian smoothed classifier.