On the Robustness of the CVPR 2018 White-Box Adversarial Example Defenses

Anish Athalye*
Massachusetts Institute of Technology
aathalye@mit.edu

Nicholas Carlini*
University of California, Berkeley
npc@berkeley.edu

Abstract

Neural networks are known to be vulnerable to adversarial examples. In this note, we evaluate the two white-box defenses that appeared at CVPR 2018 and find they are ineffective: when applying existing techniques, we can reduce the accuracy of the defended models to 0%.

1. Introduction

Training neural networks so they will be robust to adversarial examples [7] is a major challenge. Two defenses that appear at CVPR 2018 attempt to address this problem: "Deflecting Adversarial Attacks with Pixel Deflection" [6] and "Defense against Adversarial Attacks Using High-Level Representation Guided Denoiser" [4].

In this note, we show these two defenses are not effective in the white-box threat model. We construct adversarial examples that reduce the classifier accuracy to 0% on the ImageNet dataset [3] when bounded by a small ℓ_∞ perturbation of $4/255$, a stricter bound than considered in the original papers. Our attacks can construct targeted adversarial examples with over 97% success.

Our methods are a direct application of existing techniques.

2. Background

We assume familiarity with neural networks, adversarial examples [7], generating strong attacks against adversarial examples [5], and computing adversarial examples for neural networks with non-differentiable layers [1]. We briefly review the key details and notation.

Adversarial examples [7] are instances x' that are very close to an instance x with respect to some distance metric (ℓ_∞ distance, in this paper), but where the classification of x' is not the same as the classification of x. Targeted adversarial examples are instances x' whose label is equal to a given target label t.

*Equal contribution

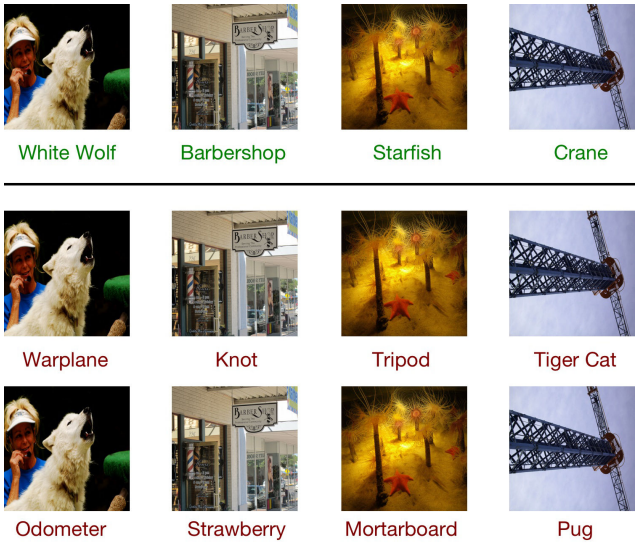

Figure 1. Original images from ImageNet validation set (row 1). Targeted adversarial examples (with randomly chosen targets) for Pixel Deflection (row 2) and High-level representation Guided Denoiser (row 3), with a ℓ_∞ perturbation of $\epsilon = 4/255$.

We examine two defenses: Pixel Deflection and High-level Representation Guided Denoiser. We are grateful to the authors of these defenses for releasing their source code and pre-trained models.

Pixel Deflection [6] proposes a non-differentiable preprocessing of inputs. Some pixels (a tunable hyperparameter) are randomly replaced with near-by pixels. This resulting image is often noisy, and to restore accuracy, a denoising operation is applied.

High-level representation Guided Denoiser (HGR) [4] proposes denoising inputs using a trained neural network before passing them to a standard classifier. This denoiser is a differentiable, non-randomized neural network. This defense has also been evaluated by Uesato et al. and found to be ineffective [8].

2.1. Methods

We evaluate these defenses under the white-box threat model. We generate adversarial examples with Projected Gradient Descent (PGD) [5] maximizing the cross-entropy loss and bounding ℓ_∞ distortion by $4/255$.

What is the right threat model to evaluate against? Many papers only claim white-box security against an attacker who is *completely unaware* the defense is being applied. HGD, for example, says "the white-box attacks defined in this paper should be called oblivious attacks according to Carlini and Wagner's definition" [4].

Unfortunately, security against oblivious attacks is not useful. We only defined this threat model in our prior work [2] to study the case of an extremely weak attacker, to show that some defenses are not even robust under this model. Furthermore, many previously published schemes already achieve security against oblivious attacks. In practice, any serious attacker would certainly consider the possibility that a defense is in place and try to circumvent it, if there is a reasonable way to do so.

Thus, security against oblivious attacks is far from sufficient to be interesting or useful in practice. Even the *black-box threat model* allows for an attacker to be aware that the defense is being applied, and only holds the exact parameters of the defense as private data. Also, our experience is that schemes that are insecure against white-box attacks also tend to be insecure against black-box attacks [2]. Accordingly, in this note, we evaluate schemes against white-box attacks.

3. Methodology

3.1. Pixel Deflection

We now show that Pixel Deflection is not robust. We analyze the defense as implemented by the authors [1]. Our evaluation code is publicly available [2].

We apply BPDA [1] to Pixel Deflection for its non-differentiable replacement operation. Our attack reduces the accuracy of the defended classifier to 0%.

In a targeted setting, we succeed with 97% probability. (Because the defense is randomized, we report success only if the image is classified as the adversarial target label 9 times out of 10.)

3.2. High-Level Representation Guided Denoiser

Next, we show that using a High-level representation Guided Denoiser is not robust in the white-box threat model. We analyze the defense as implemented by the authors [3]. Our evaluation code is publicly available [4].

We apply PGD [5] end-to-end with no modification. It reduces the accuracy of the defended classifier to 0% and achieves 100% success at generating targeted adversarial examples.

[1] https://github.com/iamaaditya/pixel-deflection

[2] https://github.com/carlini/pixel-deflection

[3] https://github.com/lfz/Guided-Denoise

[4] https://github.com/anishathalye/Guided-Denoise

4. Conclusion

As this note demonstrates, Pixel Deflection and High-level representation Guided Denoiser (HGD) are not robust to adversarial examples.

Acknowledgements

We are grateful to Aleksander Madry and David Wagner for comments on an early draft of this paper.

We thank Aaditya Prakash and Fangzhou Liao for discussing their defenses with us, and we thank the authors of both papers for releasing source code and pre-trained models.

References

[1] A. Athalye, N. Carlini, and D. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420*, 2018.

[2] N. Carlini and D. Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. *AISec*, 2017.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE, 2009.

[4] F. Liao, M. Liang, Y. Dong, T. Pang, J. Zhu, and X. Hu. Defense against adversarial attacks using high-level representation guided denoiser. In *CVPR*, 2018.

[5] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations*, 2018.

[6] A. Prakash, N. Moran, S. Garber, A. DiLillo, and J. Storer. Deflecting adversarial attacks with pixel deflection. In *CVPR*, 2018.

[7] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. *ICLR*, 2013.

[8] J. Uesato, B. O'Donoghue, A. v. d. Oord, and P. Kohli. Adversarial risk and the dangers of evaluating against weak attacks. *arXiv preprint arXiv:1802.05666*, 2018.