MagNet and “Efficient Defenses Against Adversarial Attacks” are Not Robust to Adversarial Examples

Nicholas Carlini David Wagner
University of California, Berkeley

Abstract

MagNet and “Efficient Defenses...” were recently proposed as a defense to adversarial examples. We find that we can construct adversarial examples that defeat these defenses with only a slight increase in distortion.

1 Introduction

It is an open question how to train neural networks so they will be robust to adversarial examples [11]. Recently, three defenses have been proposed to make neural networks robust to adversarial examples:

- MagNet [8] was proposed as an approach to make neural networks robust against adversarial examples through two complementary approaches: adversarial examples near the data manifold are *reformed* to lie on the data manifold that are classified correctly, whereas adversarial examples far away from the data manifold are *detected* and rejected before classification. MagNet does not argue robustness in the white-box setting; rather, the authors argue that MagNet is robust in the grey-box setting where the adversary is aware the defense is in place, knows the parameters of the base classifier, but *not* the parameters of the defense.

- An efficient defense [12] was proposed to make neural networks more robust against adversarial examples by performing Gaussian data augmentation during training, and using the BReLU activation function. The authors do not claim perfect security, but claim this makes attacks visually detectable.

- Adversarial Perturbation Elimination GAN (APE-GAN) [10] is similar to MagNet, only adversarial examples are projected onto the data manifold using a Generative Adversarial Network (GAN) [2]. We did not set out to bypass this defense, but found it to be very similar to MagNet and so we analyze it too.

In this short paper, we demonstrate these three defenses are not effective on the MNIST [5] and CIFAR-10 [4] datasets. We show that we are able to bypass MagNet with greater than 99% success, and the latter two with 100%, with only a slight increase in distortion. [1]

We defeat MagNet by making use of the *transferability* [11] property of adversarial examples: the adversary trains their own copy of the defense, constructs adversarial examples on their model, and supplies these adversarial examples to the defender. It turns out that these examples will also fool the defender’s model.

We defeat “Efficient Defenses Against Adversarial Attack” and APE-GAN by showing that existing attack can defeat them with 100% success without modification. Adversarial examples are not more visually detectable than an undefended network.

2 Background

We assume familiarity with neural networks [5], adversarial examples [11], transferability [6], generating strong attacks against adversarial examples [1] and MagNet [8]. We briefly review the key details and notation.

Notation Let $F(x) = y$ be a neural network used for classification outputting a probability distribution. Call the second-to-last layer (the layer before the the softmax layer) Z, so that $F(x) = \text{softmax}(Z(x))$. Each output y_i corresponds to the predicted probability that the object x is labelled as class i. Let $C(x) = \arg\max_i \; F(x_i)$ correspond to the classification of x on F. In this paper we are concerned with neural networks used to classify images (on MNIST and CIFAR-10).

Adversarial examples [11] are instances x' that are very close to a normal instance x with respect to some distance metric (L_2 distance, in this paper), but where $C(x') = t$ for any target t chosen by the adversary.

[1] We apologize to the reader for the seemingly incohesive format of this paper; the arXiv moderators denied separate papers as “insufficiently substantive to stand alone” and required us to submit the three independent breaks in one paper.

We generate adversarial examples with Carlini and Wagner's L_2 attack algorithm [1]. Specifically, we solve

$$\text{minimize } \|x' - x\|_2^2 + c \cdot \ell(x')$$

where the loss function ℓ is defined as

$$\ell(x') = \max(\max\{Z(x')_i : i \neq t\} - Z(x')_t, 0).$$

and the constant c is chosen via binary search.

Perhaps the most surprising property of adversarial examples is *transferability* [11]: given two different models that perform the same task, trained on different datasets, adversarial examples constructed to fool one one model often fool the other as well. This effect can be amplified [6] by constructing the adversarial examples to fool multiple local models before appling it to the target model.

MagNet [8] was recently proposed as a defense to adversarial examples under the *grey box* threat model: here, the adversary is aware that MagNet is being used to secure a given classifier, as well as the parameters of the classifier, but is not aware of the parameters of the defensive models. MagNet has two defense components:

1. The *detector* examines inputs and decides if they are adversarial or benign. There can be multiple detectors, in which case if any report "adversarial" the input is rejected.

2. The *reformer* takes inputs classified as benign by the detector(s) and reforms them to remove any remaining minor adversarial nature to the instances. The defender trains multiple reformers and selects one at random, to prevent the adversary from learning its parameters.

The authors argue these defenses complement each other: the detector should be able to detect any adversarial examples with large distortion and are far away from the true data manifold. The reformer takes instances which are close to the data manifold and maps them directly on to the manifold to remove any minor adversarial purterbations that may not have been detected.

The specific instantiation of the detector and reformer suggested by the authors are based on autoencoders. Given an autoencoder AE, to detect if an input is adversarial, compare the mean squared error $\|x - AE(x)\|_2$ with a threshold τ chosen to minimize the false positive rate. On CIFAR, the authors suggest one further detector: compute the Jensen-Shannon divergence $JSD(F(x)/T, F(AE(x))/T)$ with temperature T (set to 10 and 40), and report "adversarial" if the difference is greater than a different threshold τ', also chosen to minimize the false positive rate.

Similarly, to reform an adversarial example run it through the autoencoder and let $x \leftarrow AE(x)$: since the autoencoder is trained to reconstruct typical instances, this operation projects it back to the data manifold.

Efficient Defenses Against Adversarial Attacks [12] works by making two modifications to standard neural networks. First, the authors propose use of the Bounded ReLU activation function, defined as $BReLU(x) = \min(\max(x, 0), 1)$ instead of standard ReLU [9] which is unbounded above. Second, instead of training on the standard training data $\{(x_i, y_i)\}_{i=1}^n$ they train on $\{(x_i + N_i, y_i)\}_{i=1}^n$ where $N_i \sim \mathcal{N}(0, \sigma^2)$ is chosen fresh for each training instance. On MNIST, $\sigma = 0.3$; for CIFAR, $\sigma = 0.05$. The authors claim that despite training on noise, it is successful on [1].

APE-GAN [10] works by constructing a pre-processing network $G(\cdot)$ trained to project both normal instances and adversarial examples back to the data manifold as done in MagNet. The network G is trained with a GAN instead of an auto-encoder. Note that unlike a standard GAN which takes as input a noise vector and must produce an output image, the generator in APE-GAN takes in an adversarial example and must make it appear non-adversarial. During training, the authors train on adversarial examples generated with the Fast Gradient Sign algorithm [3]; despite this, the authors claim robustness on a wide range of attacks (including [1]).

Defense Models. We take the MagNet implementation from the authors' open-source code [2] and train our own models. Since the provided code does not include an implementation of the CIFAR defense and classifier, we implement it as described in the paper.

We were unable to obtain source code for "Efficient Defenses...". We therefore re-implement the proposed defense based on the description in the paper. We take the APE-GAN implementation from the authors open-source code [3].

We use the C&W attack given by the authors. We also release our code [4] [5] [6].

[2] https://github.com/Trevillie/MagNet

[3] https://github.com/shenqixiaojiang/APE-GAN

[4] https://github.com/carlini/MagNet

[5] https://github.com/carlini/breaking_efficient_defenses

[6] https://github.com/carlini/APE-GAN

Dataset	Model	Success	Distortion (L_2)
MNIST	Unsecured	100%	1.64
	MagNet	99%	2.25
CIFAR	Unsecured	100%	0.30
	MagNet	100%	0.45

Table 1: Mean distance to the nearest targeted adversarial example, across the first 1000 test instances, with the target chosen uniformly at random from the incorrect classes.

3 Breaking MagNet

We demonstrate MagNet is not robust by generating transferable adversarial examples with Carlini and Wagner's L_2 attack. To do this, we (as the adversary) train 32 local autoencoders to use as detectors and reformers (different than the 16 the defender will use). To generate adversarial examples, we modify the optimization formulation to the following, where reformers are denoted R_j and the detectors D_j:

$$\text{minimize } \|x' - x\|_2^2 + c \cdot \ell_c(x') + d \cdot \ell_d(x')$$

where the *classifier* loss function ℓ_c is defined as

$$\ell_c(x') = \sum_j \max(\max_{i \neq t} Z(R_j(x'))_i - Z(R_j(x'))_t, -\kappa)$$

and the *detector* loss function ℓ_d is defined as

$$\ell_d(x') = \sum_j \max(D_j(x') - \tau_j, 0).$$

Here, c and d are chosen via binary search simultaneously. The threshold τ_j is chosen such that 1% of validation images are detected as adversarial. (The true defender sets their remote τ to reject only 0.1% of instances [8]: ours is more strict to help the attack construct adversarial examples that fool the remote detectors more often.) We set the confidence $\kappa = 1$ so that the adversarial examples are more likely to transfer.

We attack by performing 10000 iterations of gradient descent with a learning rate of 10^{-2}. We did not perform hyperparameter search (e.g., picking 32 autoencoders, $\kappa = 1$, $\tau_j = 0.01$); improved search over these parameters would yield lower distortion adversarial examples.

Figures 1 and 2 contain images of targeted adversarial examples on on the secured network, and Table 1 the mean distortion required across the first 1000 instances of the test set with targets chosen uniformly at random among the incorrect classes.

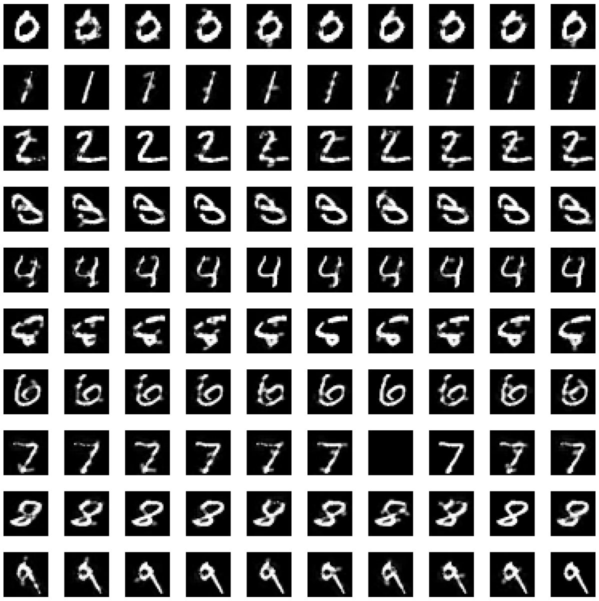


Figure 1: MagNet targeted adversarial examples for each source/target pair of images on MNIST. We achieve a 99% grey-box success (the 7 → 6 attack failed to transfer).

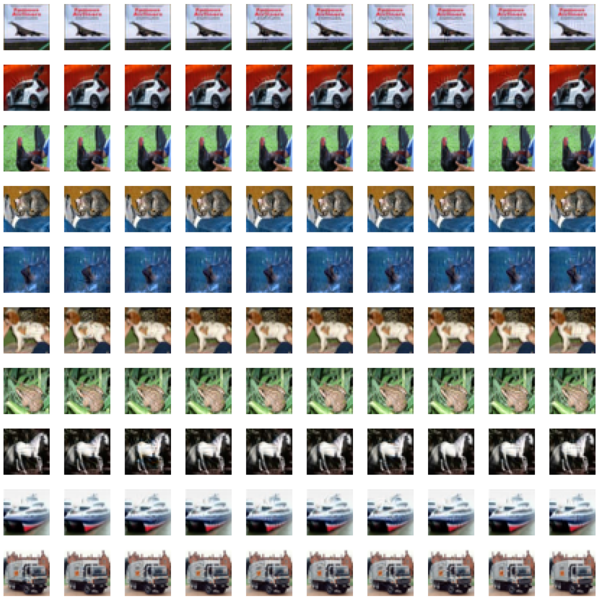


Figure 2: MagNet targeted adversarial examples for each source/target pair of images on CIFAR. We achieve a 100% grey-box success.

Dataset	Model	Distortion (L_2)
MNIST	Unsecured	2.04
	BReLU	2.14
	Gaussian Noise	2.66
	Gaussian Noise + BReLU	2.58
CIFAR	Unsecured	0.56
	BReLU	0.58
	Gaussian Noise	0.66
	Gaussian Noise + BReLU	0.67

Table 2: Neither adding Gaussian data augmentation during training nor using the BReLU activation significantly increases robustness to adversarial examples on the MNIST or CIFAR-10 datasets; success rate is always 100%.

4 Breaking "Efficient Defenses..."

We demonstrate this defense is not robust by generating adversarial examples with Carlini and Wagner's L_2 attack. We do nothing more than apply the attack to the defended network.

Figure 3 contains images of adversarial examples on the secured network, and Table 2 the mean distortion required across the first 1000 instances of the test set with targets chosen at random among the incorrect classes.

On MNIST, the full defense increases mean distance to the nearest adversarial example by 30%, and on CIFAR by 20%. This is in contrast with other forms of retraining, such as adversarial retraining [7], which increase distortion by a significantly larger amount. Interestingly, we find that BReLU provides some increase in distortion when trained without Gaussian augmentation, but when trained with it, does not help.

5 Breaking APE-GAN

We demonstrate APE-GAN is not robust by generating adversarial examples with Carlini and Wagner's L_2 attack. We do nothing more than apply the attack to defended network. That is, we change the loss function to account for the fact that the manifold-projection is done before classification. Specifically, we let

$$\ell(x') = \max(\max\{Z(G(x'))_i : i \neq t\} - Z(G(x'))_t, 0)$$

and solve the same minimization formulation.

Figure 4 contains images of adversarial examples on APE-GAN, and Table 3 the mean distortion required across the first 1000 instances of the test set with targets chosen at random among the incorrect classes.

Dataset	Model	Success	Distortion (L_2)
MNIST	Unsecured	100%	2.04
	APE-GAN	100%	2.17
CIFAR	Unsecured	100%	0.43
	APE-GAN	100%	0.72

Table 3: APE-GAN does not significantly increase robustness to adversarial examples on the MNIST or CIFAR-10 datasets.

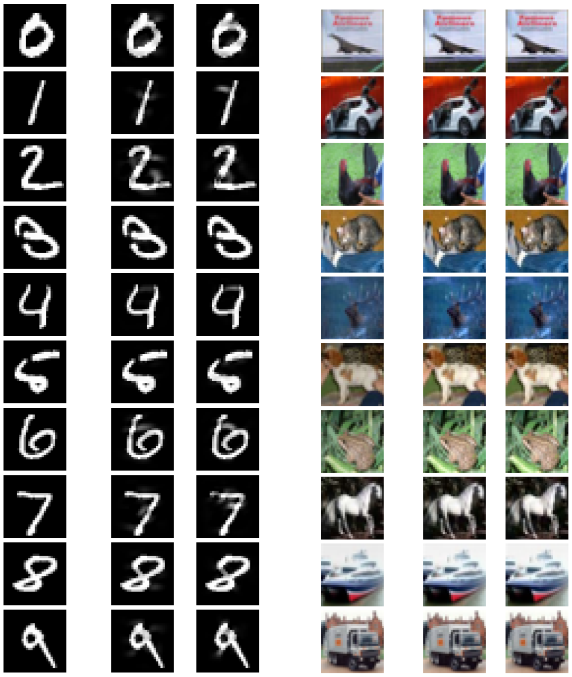
Investigating APE-GAN's Failure. Why are we able to fool APE-GAN? We compare (a) the mean distance between the original inputs and the adversarial examples, and (b) the mean distance between the original inputs and the recovered adversarial examples. We find that the recovered adversarial examples are *less similar* to the original than the adversarial examples. Specifically, the mean distortion between the adversarial examples and the original instances is 4.3, whereas the mean distortion between the recovered instances and and original instances is 5.8.

This indicates that what our adversarial examples have done is fool the generator G into giving reconstructions that are even less similar from the original than the adversarial example. This effect can be observed in Figure 4: faint lines introduced become more pronounced after reconstruction.

6 Conclusion

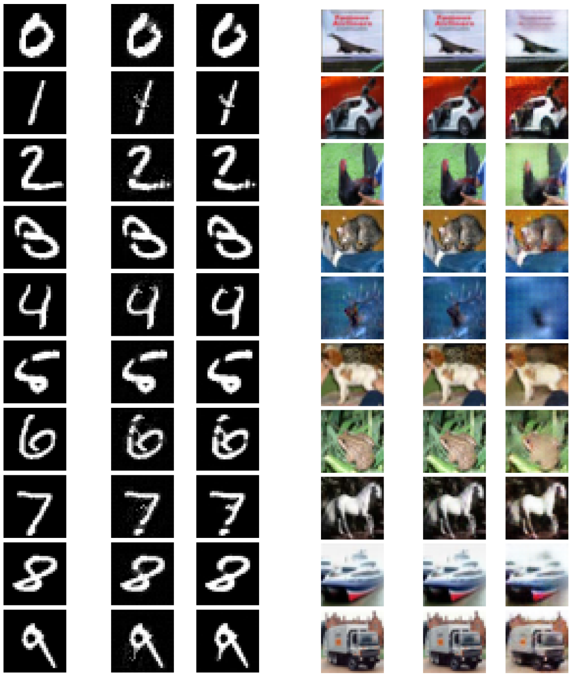
As this short paper demonstrates, MagNet is not robust to transferable adversarial examples, and combining Gaussian data augmentation and BReLU activations does not significantly increase the robustness of a neural network against strong iterative attacks. Surprisingly, we found that while all three defenses take different approaches to increasing the robustness against adversarial examples, they all give approximately the same increase in robustness ($\sim 30\%$).

We recommend that researchers who propose defenses attempt adaptive white-box attacks against their schemes before claiming robustness. Or, if arguing in the grey-box setting (even without white-box access to a given defense) it is still possible to perform an adaptive attack that succeeds with significantly higher accuracy than just generating adversarial examples against an undefended network: the adversary should generate transferable adversarial examples against *that specific defense*. We similarly recommend researchers who argue robustness under the grey-box threat model attempt similar attacks.



(a) *(b)* *(c)* *(a)* *(b)* *(c)*

Figure 3: Attacks on "Efficient Defenses..." on MNIST and CIFAR-10: *(a)* original reference image; *(b)* adversarial example on the defense with only BReLU; *(c)* adversarial example on the complete defense with Gaussian noise and BReLU.



(a) *(b)* *(c)* *(a)* *(b)* *(c)*

Figure 4: Attacks on APE-GAN on MNIST and CIFAR-10: *(a)* original reference image; *(b)* adversarial example on APE-GAN; *(c)* reconstructed adversarial example.

References

[1] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. *IEEE Symposium on Security and Privacy*, 2017.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[3] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[4] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. 2009.

[5] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[6] Y. Liu, X. Chen, C. Liu, and D. Song. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*, 2016.

[7] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[8] D. Meng and H. Chen. MagNet: a two-pronged defense against adversarial examples. In *ACM Conference on Computer and Communications Security (CCS)*, 2017. arXiv preprint arXiv:1705.09064.

[9] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 807–814, 2010.

[10] S. Shen, G. Jin, K. Gao, and Y. Zhang. APE-GAN: Adversarial Perturbation Elimination with GAN. *arXiv preprint arXiv:1707.05474*, 2017.

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

[12] V. Zantedeschi, M.-I. Nicolae, and A. Rawat. Efficient defenses against adversarial attacks. *arXiv preprint arXiv:1707.06728*, 2017.