Robust Deep Learning

Anish Madan, Dr. Saket Anand

Indraprastha Institute of Information Technology, Delhi

Abstract

Machine Learning models are deployed in various tasks including image classification, malware detection, network intrusion detection, etc. But recent work has demonstrated that even state-of-the-art deep neural networks, which excel at such tasks, are vulnerable to a class of malicious inputs known as ***Adversarial Examples****. These examples are non-random inputs that are almost indistinguishable from natural data and yet are classified incorrectly. In our work, we try to explore existence of adversarial examples, discuss some of the various attacks developed to exploit the weaknesses of deep neural networks over the years, and provide an analysis of such attacks over a subset of visually distinct classes of ImageNet.*

1. Introduction

RECENT BREAKTHROUGHS in Computer Vision have led to near human level performance in various tasks like Image Classification and Object Detection. It turns out that these "highly accurate classifiers" show an inherent weakness known as adversarial examples. ***Szegedy et al.[1]*** first noticed the existence of adversarial examples in the image classification domain, when he observed that on applying an imperceptible non-random perturbation to a test image, it was possible to change the network's prediction arbitrarily. These perturbations were found by optimizing the input to maximize the prediction error and were termed as **Adversarial Examples**.

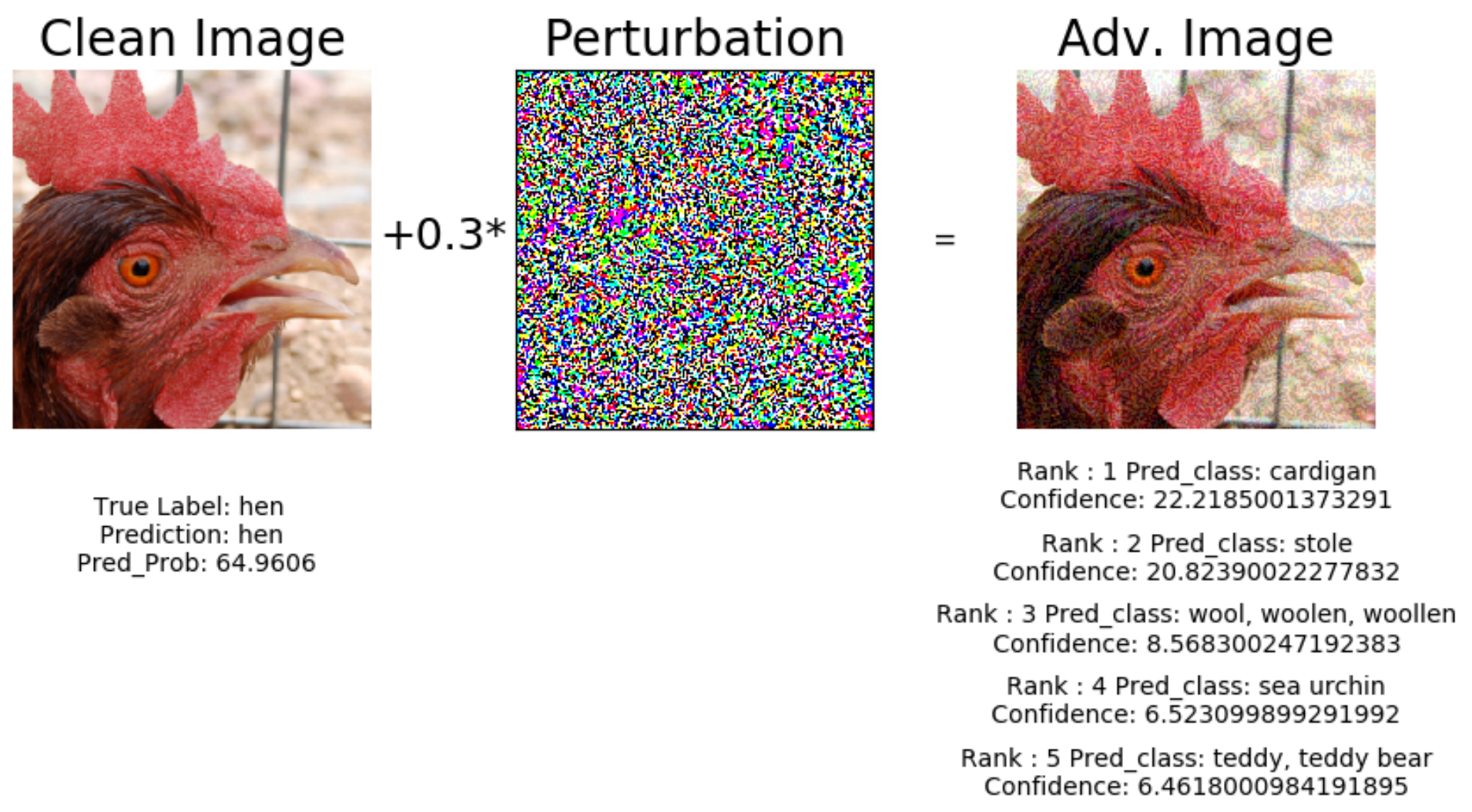
Clean Image
Perturbation
Adv. Image
+0.3*
=
True Label: hen
Prediction: hen
Pred_Prob: 64.9606
Rank : 1 Pred_class: cardigan
Confidence: 22.2185001373291
Rank : 2 Pred_class: stole
Confidence: 20.82390022277832
Rank : 3 Pred_class: wool, woolen, woollen
Confidence: 8.568300247192383
Rank : 4 Pred_class: sea urchin
Confidence: 6.523099899291992
Rank : 5 Pred_class: teddy, teddy bear
Confidence: 6.4618000984191895

Figure 2: FGSM performed on an image of hen as predicted by the unperturbed model. After adding some perturbation, we can see that the image is still a hen but our model does not predict it as hen even in its top five predictions.

Machine learning algorithms are usually designed under the assumption that models are trained on samples drawn from a distribution that is representative of test samples for which we would later make predictions[2]. However, this is not true in the case of adversarial examples.

The existence of adversarial examples is a manifestation of the difference between the real distribution D_{real}^{Ci} and the learned training distribution D_{train}^{Ci}, i.e the adversary aims to find a sample from D_{real}^{Ci} whose behavior is not captured by the learned distribution D_{train}^{Ci}. But the adversary does not know the real distribution (if we would, then no need to learn anything) so it takes a sample from D_{train}^{Ci} and tries to perturb it to craft adversarial examples. Each such example made would belong to D_{adv}^{Ci} for the class. Since the perturbations applied are tiny in nature, we know that D_{adv}^{Ci} is consistent with D_{real}^{Ci} (since a sample in D_{adv}^{Ci} would also be in D_{real}^{Ci}) , but D_{adv}^{Ci} differs from D_{train}^{Ci}.

2. Methods

2.1 Fast Gradient Sign Method

Adversarial examples not only exist in big neural networks but also for shallow linear models.
Maximum change per pixel is denoted by i.e $||\eta||_\infty < \epsilon$ where ϵ is small enough to be discarded by the camera sensor.
For linear models, we calculate the inner product between the weights of the network w and the adversarial example $\tilde{x}$, where w is the unperturbed image :

$$w^T\tilde{x} = w^Tx + w^T\eta \tag{1}$$

This equation tells us that the activation increases by $w^T\eta$. Our goal is to increase this activation as much as possible, while having capped the maximum change per pixel. This can be achieved by assigning $\eta = sign(w)$.

Let θ be the parameters of the model, x is the input image, y is the true label for x, and $J(\theta, x, y)$ is the loss function for the network. We now linearize the loss function at the current value of θ to obtain a perturbation of :

$$\eta = \epsilon sign(\nabla_x J(\theta, x, y)) \tag{2}$$

This method is known as the **Fast Gradient Sign Method[3]**.

2.2 Basic Iterative Method

The Basic Iterative Method[4] is simply an extension of the Fast Gradient Sign Method. We take multiple smaller steps of size α instead of one step ϵ like in FGSM.

$$X_0^{adv} = X \tag{3}$$

$$X_{N+1}^{adv} = Clip_{X,\epsilon}\{X_N^{adv} + \alpha sign(\nabla_X J(X_N^{adv}, y_{true}))\} \tag{4}$$

Also the image pixel values are clipped at every iteration to ensure that they are in ϵ-neighbourhood of the original image.

2.3 Carlini Wagner's L2 Attack

Carlini and Wagner(CW)[6] proposed a set of attacks based on L0, L2, $L\infty$ metric which focused only on targeted attacks since they are much more powerful than the untargeted ones. Let us now discuss CW's approach by defining the problem.
The problem formulation is similar to that defined by Szegedy et al.[1]:

minimize $||\delta||_p$ + c.$f(x+\delta)$
such that $x + \delta \in [0,1]^n$

1-D optimization techniques like binary search is used to find optimum c. Also we need to take care of the range of $x_i + \delta_i$, i.e the pixels should remain between 0 and 1. We now use the equations discussed to formulate our L2 attack. For a given datapoint x and target class t (different from class of x),we try to find w which does the following:

minimize $||0.5(tanh(w)+1) - x||_2^2 + c.f(0.5(tanh(w)+1))$

with f defined as

$$f(x') = max(max\{Z(x')_i : i \neq t\} - Z(x')_t)$$

where Z(x) gives the logits of x.

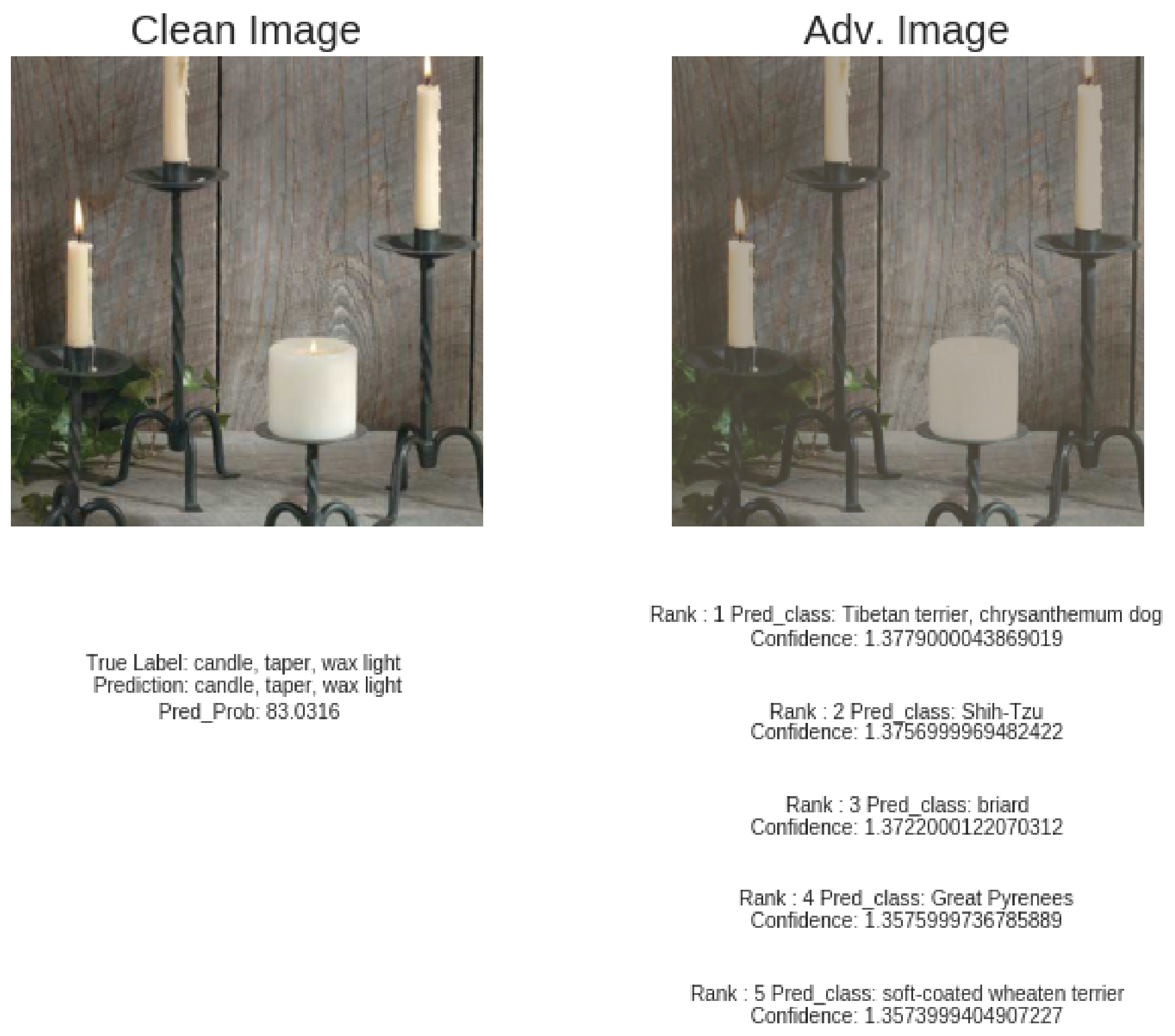
Clean Image
Adv. Image
True Label: candle, taper, wax light
Prediction: candle, taper, wax light
Pred_Prob: 83.0316
Rank : 1 Pred_class: Tibetan terrier, chrysanthemum dog
Confidence: 1.3779000043869019
Rank : 2 Pred_class: Shih-Tzu
Confidence: 1.3756999969482422
Rank : 3 Pred_class: briard
Confidence: 1.3722000122070312
Rank : 4 Pred_class: Great Pyrenees
Confidence: 1.3575999736785889
Rank : 5 Pred_class: soft-coated wheaten terrier
Confidence: 1.3573999404907227

Figure 4: CW-L2 attack on an image of a candle with the target label **Tibetan terrier**. The attack successfully changes the original label to the target class without much distortion.

2.4 Madry's PGD Attack

Madry et al.[5] provided a way to understand adversarial robustness of classifiers through the eyes of optimization. A natural **saddle point(min-max)** formulation is used to study security against adversarial attacks. This formulation helps us tie the attacks and defenses into one framework: in particular the different attacks correspond to solving the constrained optimization problem.

Consider a classifier with training data distribution **D** and with training samples as tuples (x,y) , where $x \in \mathbb{R}^d$ and $y \in [k]$ are the corresponding labels. Let us define the loss function as $L(\theta, x, y)$ and θ are the weights of the network (classifier). Empirical Risk Minimization(ERM) does not produce adversarially robust models, so we will try to change our problem to solve. We directly include the perturbations into the loss function which gives rise to the following saddle point problem:

$$\min_{\theta} \rho(\theta), where\ \ \rho(\theta) = \mathbb{E}_{(x,y)\sim D}[\max_{\delta \in S} L(\theta, x+\delta, y)] \tag{5}$$

The above equation is a form of min-max equation and consists of an inner maximization and outer minimization problem. While the inner maximization corresponds to an adversary seeking high loss, the outer minimization problem represents minimizing adversarial loss while finding corresponding weights.

3. Experiments and Results

In this section we will compare the various methods discussed and explore the behaviour of each attack. We have selected a VGG11 model which is pretrained on ImageNet. To analyze these attacks, we have selected 30 classes from ImageNet which are visually distinct from one another. For each class we choose 5 images randomly, so are test set now consists of 150 images. The images are chosen so as they classify correctly without adversaries.

3.1 Untargeted Attacks

Here we compare the untargeted attacks including FGSM, BIM, PGD.

Untargeted Attack for $\epsilon = 0.1$			
Method	No. of Steps	Top-1 Accuracy(%)	Top-5 Accuracy(%)
Natural	-	100	100
FGSM	1	2.6	19.3
BIM	1	2.6	26
PGD	6	30	72.6

Table 1: This table reports the various accuracies for different untargeted methods while fixing $\epsilon = 0.1$

Untargeted Attack for $\epsilon = 0.6$			
Method	No. of Steps	Top-1 Accuracy(%)	Top-5 Accuracy(%)
Natural	-	100	100
FGSM	1	2	7.9
BIM	2	0.6	10.6
PGD	6	12.6	33.9

Table 2: This table reports the various accuracies for different untargeted methods while fixing $\epsilon = 0.6$

3.2 Targeted Methods

Here we compare FGSM(targeted) and BIM(targeted) methods.

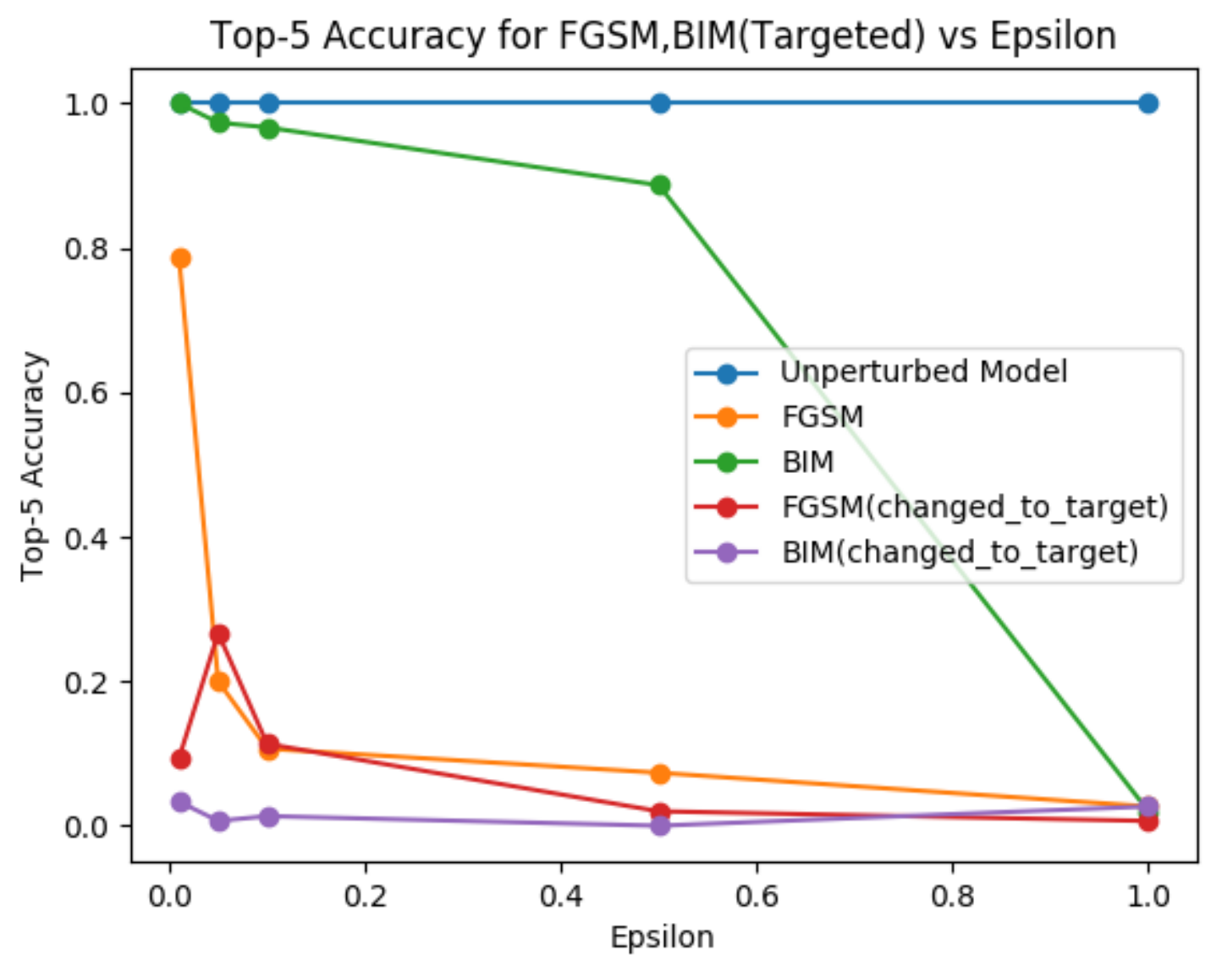

Figure 6: Comparing top-5 accuracy values for different epsilon values among various targeted methods.

3.3 Carlini Wagner L2 Attack

We discuss this attack separately since it is quite different from the other attacks. The results produced by this method are unmatched to other targeted methods.

Targeted CW L2 attack		
Metric	Top-1	Top-5
Adv target acc	100%	100%
Accuracy	100%	51.73%

Table 3: This table reports the results obtained on applying CW -L2 attack on the ImageNet subset. Adv target acc refers to the accuracy with which original labels were successfully changed to the given target label. Misclassifications report the proportion of examples misclassified by the model on attack.

References

[1] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. ICLR, abs/1312.6199, 2014. URL `http://arxiv.org/abs/1312.6199`

[2] Kathrin Grosse, Praveen Manoharan, Nicolas Papernot, Michael Backes, and Patrick McDaniel. 2017. On the (Statistical) Detection of Adversarial Examples. arXiv preprint arXiv:1702.06280 (2017).

[3] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. CoRR, abs/1412.6572, 2014. URL `http://arxiv.org/abs/1412.6572`.

[4] Alex Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. Technical report, arXiv, 2016. URL `https://arxiv.org/abs/1607.02533`.

[5] 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.

[6] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. arXiv preprint arXiv:1608.04644, 2016.