Evaluating the Effectiveness of Pixel Deflection Defending against Adversarial Attacks

Ruishan Li[1,a)] Yaokai Feng[2,b)] Kouichi Sakurai[2,c)]

Abstract: It has been shown that DNNs can be easily fooled by adding carefully crafted adversarial perturbations to input images. To defend against these vulnerabilities, many approaches that attempt to improve the robustness of DNNs have been proposed. Prakash et al. present a method called Pixel Deflection, which replacing pixel with its neighborhood in an image so that classification accuracy is significantly preserved. This is because image classifiers tend to be robust to natural noise but adversarial attacks do not. A subsequent wavelet denoising operation is used to soften the corruption from attack perturbations and pixel deflection.Both pixel deflection and wavelet denoising are transformation methods to defend against adversarial examples, the contribution of each part has not been elucidated. In this study, we evaluate the effectiveness of Pixel Deflection with five kinds of attacks in detail. The results show that that recovered accuracy is mainly from its denoising processing, wavelet denoising, not pixel deflection. Comparing pixel deflection and wavelet denoising separately with totally 5000 adversarial examples, wavelet denoising shows a bit of drop of accuracy than two together while pixel deflection does not improve accuracy obviously than no defense.

Keywords: Adversarial Attack, Adversarial Defnese, Deep Learning

1. Introduction

It is said that deep learning has been providing major breakthroughs in solving many application problems and hard scientific problems at an unprecedented scale [1], [12]. For instances, reconstruction of brain circuits; analysis of mutations in DNA; and analyzing the particle accelerator data. Deep neural networks have also become the preferred choice to solve many challenging tasks in speech recognition and natural language understanding [1]. Deep learning also became the center of attention in the field of Computer Vision. Since 2012, the Computer Vision community has made big contributions to deep learning research, enabling it to provide methods for the problems in medical science and in mobile applications. The breakthrough in AI (Artificial Intelligence) in AlphaGo Zero [25] also was originally proposed for the task of image recognition.

With the help of the continuous improvements of deep neural network models and efficient deep learning software libraries, deep learning has entered into safety and security critical applications, e.g. self-driving cars, surveillance, maleware detection, drones and robotics, and voice command recognition and facial recognition ATM and Face ID security on mobile phones. Thus, we can say that deep learning solutions are about to play a major role in our daily lives.

However, an intriguing weakness of deep neural networks in the context of image classification was discovered in the work [27]. That work showed that despite their high accuracies, modern deep networks are surprisingly susceptible to adversarial attacks with small perturbations to images that remain imperceptible to human vision system. These malicious adversarial attack images are normally carfted using an optimization procedure to search for small but effective perturbations, we will discribe them in Chapter 2. Adversarial attacks can make a neural network classifier completely change its prediction about the changed images. Moreover, the same image perturbation can fool multiple network classifiers [1]. The profound implications of these results have drawn a wide interest of researchers in adversarial attacks.

Many interesting research results have been published regarding adversarial attacks on deep learning in Computer Vision. For example, in addition to the image-specific adversarial perturbations [27], the work [16] showed the existence of universal perturbations that can fool a network classifier on any image. The work [2] showed that even 3-D print real-world objects can fool deep neural network classifiers. The work [1] presented a comprehensive survey on adversarial attacks on deep learning in Computer Vision.

Image classification CNNs (Convolutional Neural Networks) have been applied for many important real-world systems. For instance, CNNs can be used by self-driving cars to identify stop signs [20]. Such systems have become targets for adversarial attacks. Recent work has shown that classifiers in such systems can be tricked by small, carefully-crafted, imperceptible perturbations to a natural image [24]. A CNN may misclassify an image into a different class by such perturbations. For example, a ”1” is recognized into a ”9” or a stop sign into a yield sign. Obviously,

[1] Department of Informatics Graduate School of Information Science and Electrical Engineering, Kyushu University
[2] Department of Informatics Faculty of Information Science and Electrical Engineering, Kyushu University
[a)] li.ruishan.869@s.kyushu-u.ac.jp
[b)] fengyk@ait.kyushu-u.ac.jp
[c)] sakurai@inf.kyushu-u.ac.jp

defending against these vulnerabilities is critical for any application systems using CNN for image recognition. Several existing works proposed defense methods that are differentiable transformations before classification. These defenses methods appear to work well at first, but attackers can easily circumvent them by "differentiating through them", i.e. by taking the gradient of a class probability with respect to an input pixel through both the CNN and the transformation [24].

In the background mentioned above, the research on the defenses of adversarial attacks for deep learning has become one of the hot topics in this field. Prakash et al. [24] has proposed a defense method with high efficiency and high level of accuracy. In this study, we investigate this method in detail. The conception will be introduced in Chapter 4 and the result will be presented in Chapter 5.

2. Adversarial Attacks

It is said that most image classification models can be fooled [8], [27]. Several proposed techniques have been proposed to generate an image that is perceptually indistinguishable from another image but is classified into different classes. If this is done when model parameters are known, the paradigm is called white-box attacks. Otherwise, called black-box attacks.

In the work [24], a brief overview of several well-known attacks is presented including 1) Fast Gradient Sign Method (FGSM); 2) Iterative Gradient Sign Method (IGSM); 3) L-BFGS; 4) Jacobian-based Saliency Map Attack (JSMA); 5) Deep Fool (DFool); 6) Carlini&Wagner (C&W); 7) Projected Gradient Descent (PGD). They are briefly explained in this chapter.

2.1 Fast Gradient Sign Method(FGSM)

FGSM [8] is a single step attack process. It uses the sign of the gradient of the loss function, ℓ , w.r.t. to the image to find the adversarial perturbation. For a given value, FGSM is defined as:

$$\hat{x} = x + \epsilon sign(\nabla \ell(F(x), x)) \tag{1}$$

This approach means that the supply of adversarial examples is continually updated, to make them be able to resist the current version of the model.

2.2 Iterative Gradient Sign Method(IGSM)

IGSM [11] is an iterative version of FGSM. After each iteration the generated image is clipped to be within a L1 neighborhood of the original and this process stops when an adversarial image has been discovered. Both FGSM and IGSM minimize the L1 norm w.r.t. to the original image. Let x0 0 = x, then after m iterations, the adversarial image is obtained by:

$$x'_{m+1} = Clip_{x,\epsilon}\{x'_m + \alpha \times sign(\nabla \ell(F(x'_m), x'_m))\} \tag{2}$$

2.3 L-BFGS

L-BFGS(Limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm) [27], [28] tries to find the adversarial input as a box-constraint minimization problem. L-BFGS optimization is used to minimize L2 distance between the image and the adversarial example while keeping a constraint on the class label for the generated image.

In general, the exact computation of distance between two images is a hard problem, so this computation is approximated using a box-constrained L-BFGS. Concretely, performing line-search to find the minimum c > 0 for which the minimizer r of the following problem satisfies f(x + r) = l.

2.4 Jacobian-based Saliency Map Attack(JSMA)

JSMA [22] estimates the saliency of each image pixel w.r.t. to the classification output, and modifies those pixels which are most salient. This is a targeted attack, and saliency is designed to find the pixel which increases the classifier's output for the target class while tending to decrease the output for other classes.

2.5 Deep Fool(DFool)

DFool [17] is an untargeted iterative attack. This method approximates the classifier as a linear decision boundary and then finds the smallest perturbation needed to cross that boundary. This attack minimizes L2 norm w.r.t. to the original image.

2.6 Carlini&Wagner (C&W)

Carlini&Wagner (C&W) [3] is one of the strongest proposed adversarial attack. C&W updates the loss function, such that it jointly minimizes L_p and a custom differentiable loss function that uses the unnormalized outputs of the classifier (logits).

2.7 Projected Gradient Descent(PGD)

Projected Gradient Descent (PGD) [14] is an iterative variant of FGSM. In each iteration, PGD follows the update rule:

$$x'_{m+1} = \prod_c lip\{FGSM(x'_m)\} \tag{3}$$

Madry et al. observe that the local maxima of the cross-entropy loss found by PGD with 10^5 random starts are distinctive, but all have similar loss values, for both normally and adversarially trained networks. Inspired by this concentration phenomena, they propose that PGD is a universal adversary among all the first-order adversaries, i.e., attacks only rely on first-order information.

3. Defenses

Defensive strategies against adversarial examples can be categorized into three kinds.

3.1 Adversarial Training

Adversarial training [8], [11], [29] is one of the most extensively investigated defenses against adversarial attacks. It defends against adversarial perturbations by training networks on adversarial images that are generated during training. Adversarial training improves the classification accuracy of the target model on adversarial examples [8], [11], [27], [29]. On some small image datasets it even improves the accuracy of clean images [8], [27] although this effect is not found on ImageNet [6] dataset. However, adversarial training is more time consuming than training on clean images only, because online adversarial

example generation needs extra computation, and it takes more epochs to fit adversarial examples [29]. These limitations hinder the usage of harder attacks in adversarial training.

3.2 Input Transformations

Preprocessing based methods process the inputs with certain transformations to remove the adversarial noise, and then send these inputs to the target model. Gu and Rigazio [10] first propose the use of denoising auto-encoders as a defense. Osadchy et al. [19] apply a set of filters to remove the adversarial noise, such as the median filter, averaging filter and Gaussian low-pass filter. Graese et al. [9] assess the defending performance on a set of preprocessing transformations on MNIST digits, including the perturbations introduced by image acquisition process, fusion of crops and binarization. Das et al. [5] preprocess images with JPEG compression to reduce the effect of adversarialnoises. Meng and Chen [15] propose a two-step defense model, which detects the adversarial input and then reformit based on the difference between the manifolds of clean and adversarial examples. Liao et al. [13] use the reconstruction error of high-level features to guide the learning of denoisers. Pixel deflection [24] in this study replaces a random pixel with its neighbour, is also belong to here.

3.3 Gradient Masking

Another family of adversarial defenses is based on the so-called gradient masking effect [21], [23], [29]. These defenses apply some regularizers or smooth labels to make the model output less sensitive to the perturbation on input. Gu and Rigazio [10] propose the deep contrastive network, which uses a layer-wise contrastive penalty term to achieveoutput invariance to input perturbation. Nayebi and Surya [18] use saturating net-works for robustness to adversarial noises. The loss function is designed to encourage the activations to be in theirsaturating regime. The basic problem with these gradientmasking approaches is that they fail to solve the vulnerability of the models to adversarial attacks, but just make theconstruction of white-box adversarial examples more difficult. These defenses still suffer from black-box attacks [21], [29]generated on other models.

4. Conception

4.1 Pixel Deflection

In past studies, random noises have proved to be a practical method to defend against adversarial examples. This is because neural networks are robust to adversarial attacks at a certain degree, with a loss of accuracy at the same time. By contrast, adversarial examples are fragile to noises as the perturbations are carefully crafted. To avoid large accuracy loss, Prakash et al. [24] introduce a new image transformation method. Choosing a pixel from an image randomly, and then replace it with another randomly selected pixel in the previous pixel's small square neighborhood. They call this process Pixel Deflection.

4.2 Wavelet Denoising

As both adversarial perturbations and pixel deflection add noises to images, methods of removing these effects, denoiser, has become necessary. JPEG compression is a well known compression method. It has been shown that noises can be removed by JPEG compression [7], but also causes signal loss. Results also show this process can recover classification accuracy on some of the adversarial examples for neural networks [5], but lose some accuracy on clean images. Previous research has established that denoiser has a certain effect on defending adversarial attacks. Soleymani et al. [26] use wavelet decomposition to defend against adversarial examples, with 3 kinds of subbands manipulation. Wavelet denoising relies on the wavelet representation of the image. In wavelet domain, Gaussian noise can be represented by small values and can be removed by setting coefficients to a threshold. Wavelet transform is a widely used process technique in image denoising [4].

Algorithm 1: Pixel deflection transform

Input : Image I, neighborhood size r
Output: Image I' of the same dimensions as I

for $i \leftarrow 0$ **to** K **do**
  Let $p_i \sim \mathcal{U}(I)$
  Let $n_i \sim \mathcal{U}(R_p^r \cap I)$
  $I'[p_i] = I[n_i]$
end

Fig. 1 Algorithm: Pixel Deflection Transform.

Fig. 2 An image shows how pixels deflect.

5. Experiment

In this chapter, we first generate adversarial examples, then use pixel deflection defend against these adversarial examples.

We performed experiments on ImageNet Validation dataset. ImageNet Validation set has 50,000 images of 1000 classes. We use ResNet-50 pretrained model from keras. The keras ResNet-50 model has a Top-1 accuracy of 74.9% on ImageNet Validation set. Amount of images in our experiments is 1000.

As about 25% of images are misclassified by the model originally, they are considered attacking successfully without any perturbation by attacking tools. However, these cases are useless for measuring the effectivveness of attack or defence because there is no difference between input and output. Therefore, we select our 1000 images from ImageNet Validation set that can be originally classified correctly by keras ResNet-50 model.

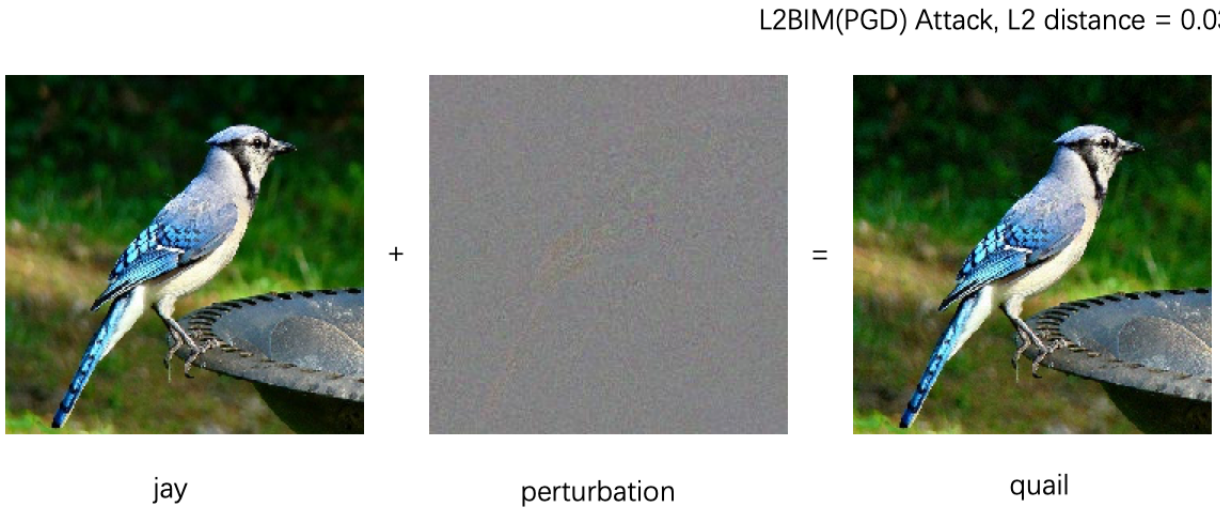

Fig. 3 An adversarial image, overlaid on a typical image, can cause a classifier to miscategorize a jay as a quail.

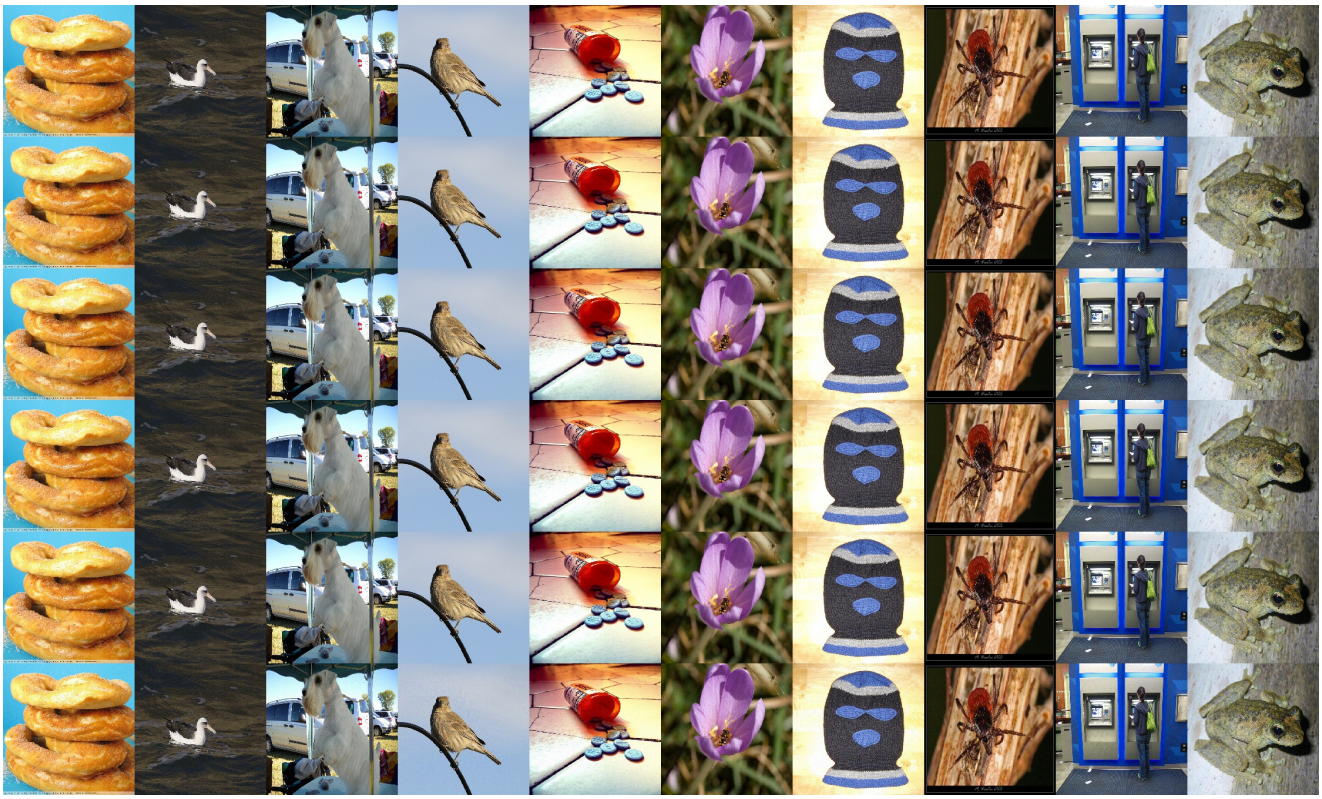

Fig. 4 Images from up to down: origin, FGSM, LBFGS, PGD, C&W, L2BIM.

5.1 Generating Adversarial Examples

We use foolbox model to generate adversarial examples. Foolbox is a Python toolbox to create adversarial examples that fool neural networks. Attacks are constricted with L-2 distance less than 0.04. We set parameter of each attack method as follows:

FGSM: confidence of miscalssification > 0.5;

L-BFGS: all default;

PGD: epsilon=0.02, stepsize=0.004, iterations=10;

C&W: confidence of miscalssification > 0.9;

L2BIM: epsilon=0.03, stepsize=0.005, iterations=20.

A sample of adversarial image is shown in Fig. 3.

Some of the adversarial images we generate. Compared with original images in Fig. 4.

Table 1 The L2 distance and classification top-1 accuracy of generated adversarial examples.

Attack	$\|L2\|$	Accuracy(%)
FGSM	0.01	8.1
L-BFGS	0.01	0.4
PGD	0.01	0
C&W	0.01	0
L2BIM	0.03	0.1

The accuracy of classfying these generated images is shown in Table 1. Python codes of generating adversarial examples are in Appendix .

5.2 Evaluating Pixel Deflection

Defence codes consist of two parts, pixel deflection and wavelet denoising. We use pixel deflection original code provided by author in [24]. Wavelet denoising in the code is called by scikit-image package. The parameters in pixel deflection are the window of size for pixel deflection and the number of pixel deflections to be performed.

We firstly verify the basic defence effect of pixel deflection against our adversarial examples.

Table 2 Top-1 accuracy on various attack before and after defense.

Attack	Before Defense(%)	After defense(%)
FGSM	8.1	86.1
L-BFGS	0.4	83.9
PGD	0	69.8
C&W	0	72
L2BIM	0.1	23.7

In Table 2, although pixel deflection doesn't get as good result as in [24], it can defend against FGSM and L-BFGS attack with accuracy of 86.1% and 83.9%. PGD and C&W attack are also be blocked by 69.8% and 72.0%. It only get bad result on L2BIM attack because of the larger L_2 distance.

But when we adjust the parameter of deflection window size and the number of deflections, we don't get the changes like in [24], the results don't show much difference. We also add and modify other parameter in pixel deflection, but still don't get much difference.

Table 3 Top-1 accuracy on various attack with pixel deflection and wavelet denoising seperately.

Attack	Only Pixel Deflection	Only Wavelet Denoising
FGSM	53.6	83.7
L-BFGS	4.0	81.6
PGD	0.29	64.1
C&W	0.12	67.1
L2BIM	0.29	18.3

To observe the changing trend of the results, we split the experiment into two parts, deflection and denoising. Without the denoising part, pixel deflection get only 53.6% accuracy on FGSM attack and amazing low accuracy near 0% on three other attacks(in Table 3). In constrast, wavelet denoising without pixel deflection gets only a bit drop comparing with the two together.

Further, we test on random noises with wavelet denoising. Result in Table shows that random noises with wavelet denoising gets even better result than pixel deflection with wavelet denoising.

Table 4 Accuracy with different numbers of deflections on L-BFGS attack.

Numbers of Deflections	Pixel deflection	Random Noises
200	4.1	28.5
1000	26.2	74.1
10000	69.8	62.9

We doubt if numbers of deflections affect the deflection, so we do another test on different numbers of deflections separately on both pixel deflection and random noises without wavelet denoising, and we get result in Table 4.

6. Conclusion

According to our results of experiments, pixel deflection and wavelet denoising together can defend low L_2 distance attacks. Based on totally 5000 adversarial examples generated by foolbox, we evaluate pixel deflection and wavelet denoising seperately. The accuracy recovered by pixel deflection and wavelet denoising is mainly from wavelet denoising, not pixel deflection.

References

[1] Akhtar, N. and Mian, A.: Threat of adversarial attacks on deep learning in computer vision: A survey, *IEEE Access*, Vol. 6, pp. 14410–14430 (2018).

[2] Athalye, A., Engstrom, L., Ilyas, A. and Kwok, K.: Synthesizing robust adversarial examples, *arXiv preprint arXiv:1707.07397* (2017).

[3] Carlini, N. and Wagner, D.: Towards evaluating the robustness of neural networks, *2017 ieee symposium on security and privacy (sp)*, IEEE, pp. 39–57 (2017).

[4] Chang, S. G., Yu, B. and Vetterli, M.: Adaptive wavelet thresholding for image denoising and compression, *IEEE transactions on image processing*, Vol. 9, No. 9, pp. 1532–1546 (2000).

[5] Das, N., Shanbhogue, M., Chen, S.-T., Hohman, F., Chen, L., Kounavis, M. E. and Chau, D. H.: Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression, *arXiv preprint arXiv:1705.02900* (2017).

[6] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K. and Fei-Fei, L.: Imagenet: A large-scale hierarchical image database, *2009 IEEE conference on computer vision and pattern recognition*, Ieee, pp. 248–255 (2009).

[7] Dziugaite, G. K., Ghahramani, Z. and Roy, D. M.: A study of the effect of jpg compression on adversarial images, *arXiv preprint arXiv:1608.00853* (2016).

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

[9] Graese, A., Rozsa, A. and Boult, T. E.: Assessing threat of adversarial examples on deep neural networks, *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, IEEE, pp. 69–74 (2016).

[10] Gu, S. and Rigazio, L.: Towards deep neural network architectures robust to adversarial examples, *arXiv preprint arXiv:1412.5068* (2014).

[11] Kurakin, A., Goodfellow, I. and Bengio, S.: Adversarial examples in the physical world, *arXiv preprint arXiv:1607.02533* (2016).

[12] LeCun, Y., Bengio, Y. and Hinton, G.: Deep learning, *nature*, Vol. 521, No. 7553, pp. 436–444 (2015).

[13] Liao, F., Liang, M., Dong, Y., Pang, T., Hu, X. and Zhu, J.: Defense against adversarial attacks using high-level representation guided denoiser, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1778–1787 (2018).

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

[15] Meng, D. and Chen, H.: Magnet: a two-pronged defense against adversarial examples, *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, pp. 135–147 (2017).

[16] Moosavi-Dezfooli, S.-M., Fawzi, A., Fawzi, O. and Frossard, P.: Universal adversarial perturbations, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1765–1773 (2017).

[17] Moosavi-Dezfooli, S.-M., Fawzi, A. and Frossard, P.: Deepfool: a simple and accurate method to fool deep neural networks, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2574–2582 (2016).

[18] Nayebi, A. and Ganguli, S.: Biologically inspired protection of deep networks from adversarial attacks, *arXiv preprint arXiv:1703.09202* (2017).

[19] Osadchy, M., Hernandez-Castro, J., Gibson, S., Dunkelman, O. and Pérez-Cabo, D.: No bot expects the DeepCAPTCHA! Introducing immutable adversarial examples, with applications to CAPTCHA generation, *IEEE Transactions on Information Forensics and Security*, Vol. 12, No. 11, pp. 2640–2653 (2017).

[20] Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B. and Swami, A.: Practical black-box attacks against deep learning systems using adversarial examples, *arXiv preprint arXiv:1602.02697*, Vol. 1,

No. 2, p. 3 (2016).

[21] Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B. and Swami, A.: Practical black-box attacks against machine learning, *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pp. 506–519 (2017).

[22] Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B. and Swami, A.: The limitations of deep learning in adversarial settings, *2016 IEEE European symposium on security and privacy (EuroS&P)*, IEEE, pp. 372–387 (2016).

[23] Papernot, N., McDaniel, P., Sinha, A. and Wellman, M.: Towards the science of security and privacy in machine learning, *arXiv preprint arXiv:1611.03814* (2016).

[24] Prakash, A., Moran, N., Garber, S., DiLillo, A. and Storer, J.: Deflecting adversarial attacks with pixel deflection, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8571–8580 (2018).

[25] Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A. et al.: Mastering the game of go without human knowledge, *Nature*, Vol. 550, No. 7676, pp. 354–359 (2017).

[26] Soleymani, S., Dabouei, A., Dawson, J. and Nasrabadi, N. M.: Defending Against Adversarial Iris Examples Using Wavelet Decomposition, *arXiv preprint arXiv:1908.03176* (2019).

[27] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. and Fergus, R.: Intriguing properties of neural networks, *arXiv preprint arXiv:1312.6199* (2013).

[28] Tabacof, P. and Valle, E.: Exploring the space of adversarial images, *2016 International Joint Conference on Neural Networks (IJCNN)*, IEEE, pp. 426–433 (2016).

[29] Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D. and McDaniel, P.: Ensemble adversarial training: Attacks and defenses, *arXiv preprint arXiv:1705.07204* (2017).