ADVERSARIAL TRAINING MAY BE A DOUBLE-EDGED SWORD

Ali Rahmati*, **Seyed-Mohsen Moosavi-Dezfooli**†, **and Huaiyu Dai***
*North Carolina State University, Raleigh, NC
†Institute for Machine Learning, ETH Zürich
{arahmat, hdai}@ncsu.edu
seyed.moosavi@inf.ethz.ch

ABSTRACT

Adversarial training has been shown as an effective approach to improve the robustness of image classifiers against white-box attacks. However, its effectiveness against black-box attacks is more nuanced. In this work, we demonstrate that some geometric consequences of adversarial training on the decision boundary of deep networks give an edge to certain types of black-box attacks. In particular, we define a metric called *robustness gain* to show that while adversarial training is an effective method to dramatically improve the robustness in white-box scenarios, it may not provide such a good robustness gain against the more realistic decision-based black-box attacks. Moreover, we show that even the minimal perturbation white-box attacks can converge faster against adversarially-trained neural networks compared to the regular ones.

1 INTRODUCTION

Adversarial examples can be crafted with intentionally designed perturbations added to the inputs to fool deep neural network classifiers. Adversarial attacks may be executed in different categories depending on the attacker's level of information, including the white-box setting (Carlini & Wagner, 2017; Goodfellow et al., 2014; Moosavi-Dezfooli et al., 2016), the score-based black-box setting (Chen et al., 2017; Ilyas et al., 2018; Narodytska & Kasiviswanathan, 2016) and the decision-based black-box setting (Chen et al., 2020; Cheng et al., 2019; Liu et al., 2019; Rahmati et al., 2020). Generally, in the white-box scenario, the effectiveness of the attack is measured by the norm of minimal perturbations required to fool the network, while in the black-box settings, norm of perturbations for a given number of queries is evaluated. Moreover, among the adversarial attacks in the literature, the most challenging ones are decision-based attacks, in which the attacker only has access to the output label of the image classifier for a given input. On the other hand, adversarial training has been shown to be an effective method to improve the robustness of the image classifiers against adversarial attacks in the white-box setting (Madry et al., 2018; Shafahi et al., 2019; Wong et al., 2019). However, the effectiveness of adversarial training is not fully understood in the more practical real-world black-box settings. Therefore, the goal of this paper is to study the effectiveness of adversarial training in improving the robustness of image classifiers against decision-based black-box attacks where the attacker's level of information from the deep neural network image classifier is the least.

In order to craft an adversarial perturbation in a decision-based black-box setting, the critical information is mostly the normal vector to the decision boundary of the image classifier. Obviously, the less information available to the attacker (which is more realistic), the less successful attack can be done. The estimation of the normal vector to the boundary in this setting is conducted with carefully designed queries on the boundary of the image classifier with the underlying assumption that the decision boundary has low mean curvature in the vicinity of inputs (Chen et al., 2020; Cheng et al., 2019; Liu et al., 2019; Rahmati et al., 2020). The goal of such black-box attacks is to reduce the number of queries as much as possible with an efficient estimate of the normal vector to the boundary. Therefore, such estimators are expected to work better if the decision boundary is less curved as most of them rely on some sort of linearization of the decision boundary. Thus, a flatter boundary would be favorable to these types of attacks.

Interestingly, it is empirically shown that adversarial training leads to neural networks with flatter decision boundaries, compared to the boundaries learned by regular training methods (Qin et al., 2019; Moosavi-Dezfooli et al., 2019). We will show that such feature of the adversarially-trained networks is indeed favorable for black-box attacks. The goal of this paper is to show some evidence that although the adversarial training improves the robustness of deep image classifiers effectively (i.e., it increases the minimum distance of datapoints to the boundary) against the minimal norm perturbation white-box attacks, it becomes less effective in more practical attack settings. That is, decision-based black-box attacks can exploit the excessive flatness caused by adversarial training. In particular, we define a metric called *robustness gain* as the ratio of ℓ_2 norm of adversarial perturbation required to fool the adversarially-trained network to that for the regular network. We observe that the robustness gain against the real-world attacks in which the least amount of information is available to the attacker is not as good as the one in the ideal white-box scenario with full information. We also show that even iterative white-box attacks may converge faster against adversarially-trained networks due to their flatter boundary and more linear behaviour of such networks.

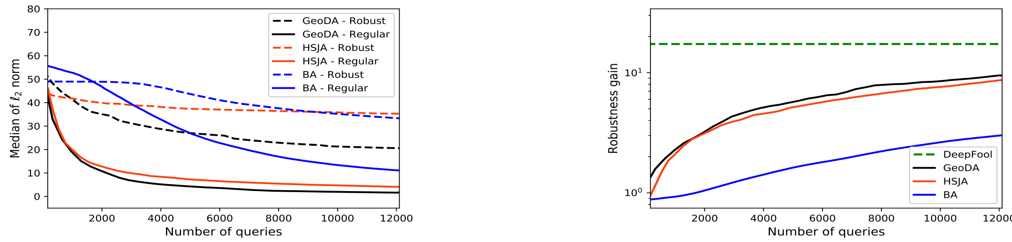
The rest of the paper is organized as follows. In Section 2, we study the effectiveness of adversarial training against decision-based black-box attacks. In Section 3, we evaluate the white-box attack performance against adversarially-trained networks. Finally, in Section 4, we conclude the paper and provide the direction forward.

2 ADVERSARIALLY-TRAINED NETWORKS AND DECISION-BASED BLACK-BOX ATTACKS

Although it is known that adversarial training is a quite effective approach to make image classifiers robust in the white-box settings, their performance in the more practical settings is not well investigated. In this section, our goal is to evaluate the effectiveness of adversarial training against decision-based black-box attacks in which the attacker has only access to the output label of the image classifier for a given input. We consider the most challenging setting in which only the top-1 label of the deep classifier is available to the attacker. Generally, such attacks aim to obtain the ℓ_2-minimized norm perturbation with the minimum number of queries.

The most successful decision-based black-box attacks in the literature (Chen et al., 2020; Cheng et al., 2019; Liu et al., 2019; Rahmati et al., 2020) aim to estimate the normal vector to the decision boundary with a local linearization of the decision boundary at a randomly obtained boundary point. Such estimators rely on the fact that the boundary of the image classifiers has a low mean curvature in the vicinity of the data samples. As a result, the less curved the boundary is, the more valid such an assumption is. An important observation regarding adversarial training (Madry et al., 2018) is that, in addition to imposing a larger distance between the data point and the decision boundary (hence resulting in a higher robustness), the decision regions of adversarially trained networks get flatter and more regular (Qin et al., 2019). In particular, the curvature of the decision boundary decreases after adversarial training (Moosavi-Dezfooli et al., 2019). It is intuitive that a normal vector estimator will provide a better estimation on smoother boundaries regardless of the boundary point at which the estimation is done. In particular, the flatter the boundary the more aligned the directions of the normal vectors at different boundary points are. In the extreme case that the boundary is a hyperplane, the normal vectors to the boundary are in the same direction throughout the hyperplane. Thus, interestingly, this results in a better estimation of the normal vector to the boundary for a adversarially-trained classifier in the black-box setting. This interesting observation might be exploited as a future direction to develop new attacks with potentially a smaller number of queries to estimate the normal vector leading to a more effective attack.

Although adversarial training provides a quite impressive robustness against attackers by increasing the distance of the input to the boundary in the white-box setting, our goal is to see how such neural networks behave in query-limited black-box settings. We define a metric called *robustness gain* $\eta = \ell_2^{\text{robust}}/\ell_2^{\text{regular}}$ as the ratio of ℓ_2 norm of the adversarial perturbation required to fool the adversarially-trained network to that for the regular network. We conduct the experiments on a pre-trained ResNet-50 (He et al., 2016) called the *regular* network and the adversarially-trained ResNet-50 (Madry et al., 2018) called the *robust* network throughout the paper. We consider a random set of 300 correctly classified images by both networks from the ILSVRC2012's validation set (Deng et al., 2009).



(a) (b)

Figure 1: (a) Performance comparison of different black-box attacks for both regular and adversarially-trained ResNet50. (b) The robustness gain for ℓ_2 norm under different attack scenarios.

Black-box attacks performance evaluation We compare the performance of black-box attacks HSJA (Chen et al., 2020), GeoDA (Rahmati et al., 2020), boundary attack (BA) (Brendel et al., 2018) on both regular and adversarially-trained ResNet-50 networks in Fig. 1a. As shown, for a given query budget, the ℓ_2 norm of perturbations for the attacks against the adversarially-trained network is larger compared to that of the regular network as expected. However, an interesting observation is that while GeoDA has almost the same ℓ_2 norm as HSJA for the regular network, it provides much smaller ℓ_2 norm for perturbations against the adversarially-trained network compared to HSJA for a fixed amount of query budget. The main reason for this phenomenon is that GeoDA is *explicitly* built based on the assumption that the boundary of the classifier has a low mean curvature. On the other hand, adversarially trained-networks has flatter decision boundaries which actually gives an edge to GeoDA. Thus, to attack adversarially-trained networks more efficiently, it is beneficial for the attackers to deploy attacks exploiting the flatness of the decision boundary.

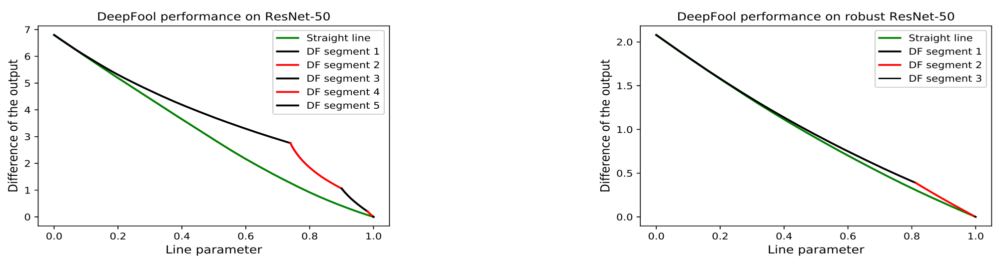
Robustness gain In this part, our goal is to evaluate how much adversarially-trained networks can improve the robustness under various kind of attacks. We plot the *robustness gain* for different attacks in Fig. 1b. The larger the η for a given attack is, the better the adversarial training can improve the robustness compared to the case of the regular network. In Fig. 1b, it is observed that η is equal to approximately 17 (see Table 1) for the white-box attack DeepFool (DF) (Moosavi-Dezfooli et al., 2016) which is a quite good improvement. However, in the case of decision-based black-box setting, first, it can be seen that the robustness gain is lower than that of the white-box attack in general. Second, by extracting more information from the image classifier (getting more queries), the robustness gain increases. It implies that the less information attacker knows about the image classifier, i.e., the more practical the attack scenario is (e.g., from white-box to black-box, or by decreasing the number of queries in the black-box setting), the less the adversarially-trained network can improve the robustness. That being said, adversarial training for the deep image classifiers is much more effective against white-box scenarios than against black-box scenarios.

3 ADVERSARIALLY-TRAINED NETWORKS AND MINIMAL NORM WHITE-BOX ATTACKS

We already discussed the effectiveness of the adversarially-trained network with respect to the number of required queries against decision-based black-box attacks in the query-limited regime. In the white-box scenario, we evaluate the effectiveness through the number of required iterations for the convergence of a minimal ℓ_2 norm perturbation white-box attack. To this end, we choose a minimal ℓ_2 norm white-box attack DeepFool (Moosavi-Dezfooli et al., 2016) and compare its performance on an adversarially-trained (Madry et al., 2018) and a regular ResNet-50 in Table 1. The main reason we choose DeepFool is its dependence on linearizing the output function of the classifier. The algorithm starts with locally linearizing the output function of the classifier and repeats such an approximation iteratively to compensate for the effect of the non-linearity of the output function. The more linear the output function of the image classifier is, the fewer iterations required for DeepFool to converge. Interestingly, despite that the adversarially-trained network requires perturbations with larger ℓ_2 norm to be fooled, due to the more linear behavior of its output function, DeepFool converges faster on this network. In this sense, one can conclude that there is a trade-off to attack adversarially-trained networks even in the white-box setting.

	Median of Iterations	**Max of Iterations**	**ℓ_2 norm**
Regular (He et al., 2016)	4	15	0.209
Adv. trained (Madry et al., 2018)	2	4	3.618

Table 1: DeepFool (Moosavi-Dezfooli et al., 2016) performance on adversarially-trained and regular ResNet-50 networks (over 1000 correctly classified samples from ImageNet validation set).



(a) (b)

Figure 2: Performance evaluation of DeepFool over different iterations on (a) Regular ResNet 50 network. (b) Adversarially-trained ResNet 50 network.

In the next experiment, the goal is to qualitatively study the behaviour of output function along the trajectory of the iterations of DeepFool for a single data point. In this case, DeepFool requires 3 and 5 iterations to converge for adversarially-trained and regular ResNet-50, respectively. We consider the difference of the logits corresponding to the original and the adversarial labels for our evaluation. We track this difference along two paths: 1) the straight path between the original image and the DeepFool adversarial example (i.e., green line in Figs. 2a and 2b), and 2) the path taken by DF in each iteration (i.e. black and red line segments). We generate images on the line from the original image to the minimal perturbation adversarial example obtained by DeepFool. By varying the line parameter t, we consider the images along the line $\boldsymbol{x} = \boldsymbol{x}_0 + t(\boldsymbol{x}_{adv} - \boldsymbol{x}_0)$, where $t = 0$ corresponds to the original image and $t = 1$ gives the adversarial image which falls on the boundary. When the image is on the clean label side, the output value of the clean label is larger than the adversarial label. Approaching the boundary, this difference decreases where on the boundary the difference is equal to zero and the transition occurs. Assuming $\boldsymbol{x}_i$ as the output of DeepFool in iteration i, each line segment i (i.e. black and red segments in Figs. 2a and 2b) is corresponding to the images on the line $\boldsymbol{x} = \boldsymbol{x}_{(i-1)} + t(\boldsymbol{x}_i - \boldsymbol{x}_{(i-1)})$ for $t \in [0, 1]$, where $\boldsymbol{x}_i = \boldsymbol{x}_{adv}$ if i is the last iteration. First, it can be seen that the straight path (green line) is much closer to the path constructed with DeepFool iterations for the adversarially-trained network compared to that of regular network. Second, it is shown that even in each line segment corresponding to each iteration traversed by DeepFool algorithm, there is more non-linearity in regular networks. As a result, it can be seen that although adversarial training improve the robustness (increases the minimal ℓ_2 norm required for successful attack), it gives an edge to the attacker (with a smaller number of iterations to converge) due to more linear behaviour of adversarially-trained networks.

4 Conclusion and Direction forward

We showed that although adversarial training is quiet effective against white-box attacks, in query-limited decision-based black-box attacks, it may not perform as efficiently as in the case for the white-box attacks. We demonstrated that since adversarial training leads to a significantly flatter boundary and a more linear behavior of the image classifier, it can give an edge to certain types of black-box attackers whose goal is to estimate the normal vector to the boundary. This feature of the adversarially-trained networks can also provide a chance for minimal norm perturbation whit-box attacks to produce adversarial examples with a smaller number of iterations. As a future direction, such properties of adversarially-trained networks can be exploited, especially in the black-box settings, to design more effective attacks against such networks.

REFERENCES

Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In *International Conference on Learning Representations*, 2018.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pp. 39–57, 2017.

Jianbo Chen, Michael I Jordan, and Martin J Wainwright. Hopskipjumpattack: A query-efficient decision-based attack. In *2020 IEEE symposium on security and privacy (sp)*, pp. 1277–1294, 2020.

Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pp. 15–26. ACM, 2017.

Minhao Cheng, Simranjit Singh, Patrick H Chen, Pin-Yu Chen, Sijia Liu, and Cho-Jui Hsieh. Sign-opt: A query-efficient hard-label adversarial attack. In *International Conference on Learning Representations*, 2019.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, 2009.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *International Conference on Machine Learning*, pp. 2137–2146. PMLR, 2018.

Yujia Liu, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. A geometry-inspired decision-based attack. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4889–4897. IEEE Computer Society, 2019.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 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, Jonathan Uesato, and Pascal Frossard. Robustness via curvature regularization, and vice versa. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9078–9086, 2019.

Nina Narodytska and Shiva Prasad Kasiviswanathan. Simple black-box adversarial perturbations for deep networks. *arXiv preprint arXiv:1612.06299*, 2016.

Chongli Qin, James Martens, Sven Gowal, Dilip Krishnan, Krishnamurthy Dvijotham, Alhussein Fawzi, Soham De, Robert Stanforth, and Pushmeet Kohli. Adversarial robustness through local linearization. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Ali Rahmati, Seyed-Mohsen Moosavi-Dezfooli, Pascal Frossard, and Huaiyu Dai. GeoDA: a geometric framework for black-box adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8446–8455, 2020.

Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! In *Advances in Neural Information Processing Systems*, pp. 3358–3369, 2019.

Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations*, 2019.