BUFFER ZONE BASED DEFENSE AGAINST ADVERSARIAL EXAMPLES IN IMAGE CLASSIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent defenses published at venues like NIPS, ICML, ICLR and CVPR are mainly focused on mitigating white-box attacks. These defenses do not properly consider adaptive adversaries. In this paper, we expand the scope of these defenses to include adaptive black-box adversaries. Based on our study of these defenses, we develop three contributions. First we propose a new metric for evaluating adversarial robustness when clean accuracy is impacted. Second, we create an enhanced adaptive black-box attack. Third and most significantly, we develop a novel defense against these adaptive black-box attacks. Our defense is based on a combination of deep neural networks and simple image transformations. While straight forward in implementation, this defense yields a unique security property which we term buffer zones. We argue that our defense based on buffer zones offers significant improvements over state-of-the-art defenses. We verify our claims through extensive experimentation. Our results encompass three adversarial models (10 different black-box attacks) on 11 defenses with two datasets (CIFAR-10 and Fashion-MNIST).

1 INTRODUCTION

There are many applications based on Convolution Neural Networks (CNNs) such as image classification (Krizhevsky et al., 2012; Simonyan & Zisserman, 2015), object detection (Girshick, 2015; Ren et al., 2015), semantic segmentation (Shelhamer et al., 2017) and visual concept discovery (Wang et al., 2017). However, it is well-known that CNNs are highly susceptible to small perturbations η which are added to *benign* input images x. As shown in (Szegedy et al., 2013; Goodfellow et al., 2014), by adding *visually imperceptible* perturbations to the original image, adversarial examples x' can be created, i.e., $x' = x + \eta$. These adversarial examples are misclassified by the CNN with high confidence. Hence, making CNNs secure against this type of attack is a significantly important task.

In general, adversarial machine learning attacks can be categorized as either white-box or black-box. This categorization depends on how much information about the classifier is necessary to run the attack. The majority of the literature has focused on white-box attacks (Tramer et al., 2020; Athalye et al., 2018; Carlini & Wagner, 2017) where the model/classifier parameters are known. Likewise, the majority of defenses have been designed with the goal of thwarting white-box attacks (Papernot et al., 2016a; Kurakin et al., 2016; Tramèr et al., 2017; Cao & Gong, 2017; Metzen et al., 2017; Feinman et al., 2017; Xie et al., 2018; Meng & Chen, 2017; Guo et al., 2017; Srisakaokul et al., 2018; Raff et al., 2019; Roth et al., 2019; Pang et al., 2019). In this paper, we focus on black-box attacks, where the classifier parameters are hidden or assumed to be secret. By analyzing defenses through a black-box adversarial lens, we complete the security picture and offer both new attack and defense perspectives to the community. Specifically we make the following contributions: We develop a new metric for understanding the efficiency of defenses. We propose a novel black-box attack called the mixed black-box attack. This new attack shows the inherent weakness of certain types of defenses. We analyze previously proposed defenses under new black-box adversaries. Lastly, based on our study of previous defenses, we develop a new more robust black-box defense which we term Buffer Zones (BUZz).

The rest of the paper is organized as follows: In section 2 we discuss black-box adversaries, why we focus on certain attacks and our new mixed black-box attack. In section 3 we discuss the concept of buffer zones as a defense technique. We also discuss related defenses from the literature that we

compare to buffer zones. In section 4 we explain how to properly understand the trade off between security and clean accuracy in defenses using a newly proposed metric. We give experimental results for all 11 defenses and 10 attacks in section 5. Lastly we offer concluding remarks in section 6.

2 ATTACKS

The general setup in adversarial machine learning for both white-box and black-box attacks is as follows (Yuan et al., 2017): We assume a trained classifier f with a correctly identified sample x with class label y. The goal of the adversary is to modify x by some amount η such that $f(x+\eta)$ produces class label $\hat{y}$. In the case of untargeted attacks, the attack is considered successful as long as $\hat{y} \neq y$. In the case of targeted attacks, the attack is only successful if $\hat{y} \neq y$ and $\hat{y} = t$ where t is a target class label specified by the adversary. For both untargeted and targeted attacks, typically the magnitude of η is limited (Goodfellow et al., 2014) so that humans can still visually recognize the image.

The difference between white-box and black-box attacks lies in how η is obtained. In white-box attacks, η may be computed through backpropagation on the classifier or by formulating the attack as an optimization problem (Szegedy et al., 2013; Carlini & Wagner, 2017; Goodfellow et al., 2015) which takes into account the classifier's trained parameters. Black-box attacks on the other hand do not have access to the classifier's parameters when generating η and must rely on other information.

In this paper, we focus on black-box adversaries which utilize adaptive attacks (Papernot et al., 2017). A natural question is why do we focus on adaptive black-box type attacks? We do so for the following three reasons:

1. Gradient masking makes it possible for a defense to give a false sense of security (Athalye et al., 2018) against white-box attacks. This means that demonstrating robustness to white-box attacks does not always imply robustness to black-box attacks. Hence, there is a need to extensively test both white and black-box attacks.
2. State-of-the-art white-box attacks on published defenses have been extensively studied in the literature (Tramer et al., 2020; Athalye et al., 2018; Carlini & Wagner, 2017). The level of attention given to adaptive black-box attacks in defense papers is significantly less. By focusing on adaptive black-box attacks, we seek to complete the security picture. This full security picture means that current defenses we analyze, have not only white-box attacks (from their own publication) but, also adaptive black-box results. Future defenses can use our framework and publicly available source code to aid in their own adaptive analysis. This completed security spectrum brings us to our third point.
3. By completing the security picture (with adaptive black-box attacks) we allow the readers to compare defense results. This comparison can be done because the same adversarial model, dataset and attack is used for each defense. This is in stark contrast to adaptive white-box attacks which may require different adversarial models and different security assumptions for each attack. As noted in (Carlini et al., 2019) it is improper to compare the robustness of two defenses under different adversarial models.

2.1 BLACK-BOX ATTACK VARIATIONS

A. Pure black-box attack (Szegedy et al., 2014; Papernot et al., 2016b; Athalye et al., 2018; Liu et al., 2017): The adversary is *only* given knowledge of a training data set $\mathcal{X}_0$.

B. Oracle based black-box attack (Papernot et al., 2017): The attacker does not have access to the original training dataset, but may generate a synthetic dataset S_0 similar to the training data. The adversary can adaptively generate synthetic data and query the defense f to obtain class labels for this data. The synthetic dataset S_0 is then used to train the synthetic model. It is important to note the adversary does not have access to the original training dataset $\mathcal{X}_0$. In this paper, we propose a new version of this attack which we call the *Mixed Black-Box attack*. In this attack the adversary is given the original training dataset, the ability to generate synthetic data and query access to the defense to label the data. In this way, the adversary can train a synthetic model whose behavior mirrors that of the defense more precisely. Experimentally, we show that this attack works better on certain types of

randomized defenses than both boundary and pure black-box attacks (Szegedy et al., 2014; Papernot et al., 2016b; Athalye et al., 2018; Liu et al., 2017; Chen et al., 2020; Chen & Gu, 2020).

C. Boundary black-box attack (Chen & Jordan, 2019): In this type of attack the adversary has query access to the classifier and only generates a single sample at a time. The main idea of the attack is to try and find the boundaries between the class regions using a binary search methodology and a gradient approximation for the points located on the boundaries.

D. Score based black-box attacks In the literature these attacks are also called Zeroth Order Optimization based black-box attacks (Chen et al., 2017). The adversary adaptively queries the defense to approximate the gradient for a given input based on a derivative-free optimization approach. This approximated gradient allows the adversary to directly work with the classifier of the defense. Another attack in this line is called SimBA (Simple Black Box Attack) (Guo et al., 2019). Unlike all the previously mentioned attacks, this attack requires the score vector $f(x)$ to mount the attack, instead of merely using the hard label.

The only type of black-box attack we do not consider in our analysis from the ones enumerated above, is the score based black-box attack. Just like white-box attacks are susceptible to gradient masking, score based black-box attacks can be neutralized by a type of gradient masking (Carlini et al., 2019). This means defenses can appear to be secure to score based black-box attacks, while actually not offering true black-box security. Furthermore, it has been noted that a decision (hard label) based black-box attack represents a more practical adversarial model (Chen et al., 2020). Therefore, we slightly focus our scope on the three other black-box variants.

We implement the pure black-box attack and mixed black-box attacks. In both these types of attacks adversarial samples are generated from the synthetic model using six different methods, FGSM (Goodfellow et al., 2014), BIM (Kurakin et al., 2017), MIM (Dong et al., 2018), PGD (Madry et al., 2018), C&W (Carlini & Wagner, 2017) and EAD (Chen et al., 2018). We also consider boundary black-box attacks. Here we implement the original boundary attack, the Hop Skip Jump Attack (HSJA) (Chen et al., 2020), as well as the newly proposed Ray Searching Attack (RayS) (Chen & Gu, 2020). In total these attacks represent fourteen different ways to generate black-box adversarial examples.

3 DEFENSES

The field of adversarial defenses is rapidly expanding. For example, in 2019 alone multiple defense papers were released almost every month[1]. To examine every proposed defense is beyond the scope of this paper. Instead, we focus our analysis on eleven recent, related and/or popular defenses. The related defenses we consider are Barrage of Random Transforms (BaRT) (Raff et al., 2019), The Odds are Odd (Odds) (Roth et al., 2019), Ensemble Diversity (ADP) (Pang et al., 2019), Madry's Adversarial Training (Madry) (Madry et al., 2018), Multi-model-based Defense (Mul-Def) (Srisakaokul et al., 2018), Countering Adversarial Images using Input Transformations (Guo) (Guo et al., 2017), Ensemble Adversarial Training: Attacks and Defenses (Tramer) (Tramèr et al., 2017), Mixed Architectures (Liu) (Liu et al., 2017), Mitigating adversarial effects through randomization (Xie) (Xie et al., 2018), Thresholding Networks (a basic proof of concept defense developed in this paper) and Buffer Zones (BUZz), the main technique proposed in this paper. In general, adversarial defenses can be divided based on several underlying defense mechanisms. While the definitions we provide here are by no means absolute, they give us a way to better understand and analyze the field.

1. **Multiple models** - The defense uses multiple classifiers for prediction. The classifier outputs may be combined through averaging (i.e. ADP), randomly picking one classifier from a selection (Mul-Def) or through majority voting (Mixed Architecture).
2. **Image transformations** -The defense applies image transformations before classification. In some cases, the transformation may be randomized (Xie and BaRT) or fixed (Guo).
3. **Adversarial training** - The classifier is trained to correctly recognize adversarial examples with their correct label. Madry, Mul-Def and Tramer all use adversarial training.

[1] https://nicholas.carlini.com/writing/2019/all-adversarial-example-papers.html

4. **Adversarial detection** - The defense outputs a null label if the sample is considered to be adversarially manipulated. Odds employs an adversarial detection mechanism, as does the vanilla thresholding network we consider as a proof of concept defense in this paper.

With so many different kinds of defenses, a natural question is why do we propose another? In short, the answer is because no current defense we analyze performs well against ALL types of black-box attacks and offers a flexible trade-off between security and clean accuracy. For example, adversarially trained networks like Madry perform poorly against pure black-box attacks. Randomized defenses like Xie and Mul-Def work well against boundary attacks but fail against mixed black-box attacks which can adapt to the randomization. If we want to increase their security, its not immediately clear how much clean accuracy will be impacted. Likewise, if we want greater clean accuracy, without completely abandoning the defense, it is not obvious how this can be accomplished. In BUZz by adding more networks this trade-off between security and clean accuracy is transparent. BUZz is also one of the only defenses that performs well across multiple types of black-box attacks.

We present full experimental results in section 5 to support these claims and give an individual analysis of every defense with respect to black-box attacks in the supplementary material. Our main focus is to create a defense where the other proposed methods fall short. We strive to create a high fidelity defense (BUZz) that provides flexibility between security and clean accuracy.

3.1 Buffer Zones

The BUZz defense is based on the concept of buffer zones. Buffer zones are the regions in between classes where if an input falls in this region, it is marked as adversarial. In theory, this forces the adversary to add noise η greater than a certain magnitude in order to overcome the buffer zone. Because an attack fails if the noise becomes visual perceptible to humans, the adversary is limited in terms of the magnitude of η. In many cases this means the adversary may not be able to overcome the buffer zone and therefore cannot fool the classifier. Buffer zones are shown both in a theoretical diagram and with actual experimental results in Figure 2. The natural question is how can buffer zones be implemented in classifiers? In this section we discuss different techniques that can be used to create buffer zones.

3.2 Realizing Buffer Zones through Multiple Networks

Buffer zones can be created through the use of multiple networks. A naïve approach to this method would be to simply use networks with different architectures. However, we show that merely using different architectures does not yield security. Specifically, we test such a defense in our results by using one VGG16 and one ResNet56 with majority voting (we denote this as the Liu defense). This has also been shown in the literature in (Liu et al., 2017). Other examples of architectural defenses not yielding security include ADP and Mul-Def (which we test in this paper). Instead to break transferability between networks we introduce secret image transformations for each classifier. Our defense composed of multiple classifiers (each with their own transformations) is depicted in Figure 1. Each CNN has two *simple unique image transformations* as shown in Figure 1. The first is a fixed randomized linear transformation $c(x) = Ax + b$, where A is a matrix and b is a vector.

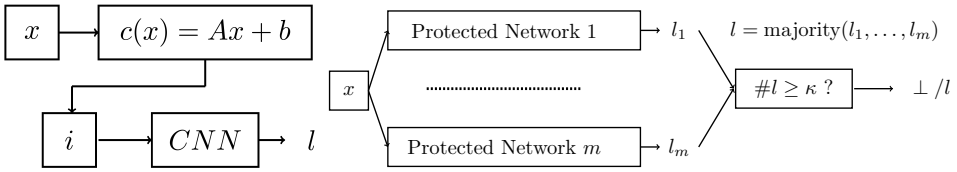

Figure 1: Design of a single network with transformations (left) in BUZz and the entire BUZz defense comprised of multiple networks each with their own transformation. The networks combined output is decided through majority voting and a threshold κ (right).

After the linear transformation a resizing operation i is applied to the image before it is fed into the CNN. The CNN corresponding to c and i is trained on clean data $\{i(c(x))\}$. Multiple CNNs are used, each with their own resizing operation and A and b components as shown in Figure 1.

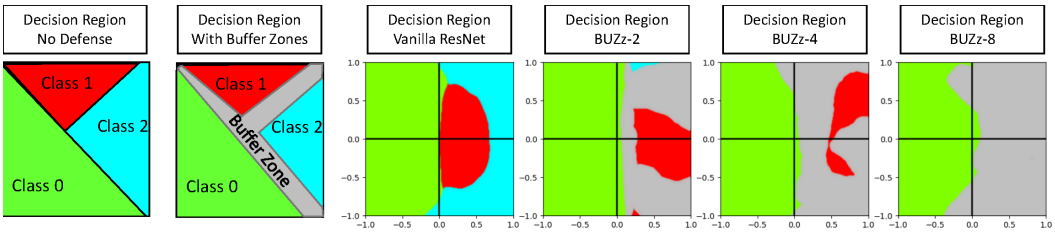

Figure 2: Decision regions with and without buffer zones $\perp$.

Security Argument. In a multi-classifier defense, buffer zones can be established by using absolute majority voting and diminishing the transferability among the different classifiers through image transformations. To decrease the transferability, we can train each classifier on a different transformed input x. From (Guo et al., 2017) we know adversarial examples are sensitive to image transformations which either distort the value of the pixels in the image or change the original spatial location of the pixels.

For any adversarial attack on BUZz, the adversary must generate a noise η for input x. Since the adversarial sample $x' = x+\eta$ is input into every network of BUZz, the j-th network will apply its own set of transformations and do classification: $i_j(c_j(x')) = i_j(c_j(x+\eta))$, which due to the linearity of $i_j(c_j(\cdot))$ is equal to $i_j(c_j(x)) + i_j(c_j(\eta))$. Therefore, networks receive different input noise $i_j(c_j(\eta)) \neq \eta$. This inherently makes the task of the adversary difficult because they must generate x' which produces the same wrong output for every classifier's transformation, $i_j(c_j(x')), j = 1..., m$. If even one classifier disagree with the majority, the sample is marked as adversarial and the attack fails. It is important to note that in this paper we experimentally established that image resizing and linear transformations can reduce transferability. However, there may be other image transformations that can also accomplish this goal.

Image Transformation Defenses:A few simple questions arise when dealing with image transformations in security. For example, can only one network with image transformations be used without retraining? We test this concept using the defense by Xie (and we show it performs worse than BUZz under the mixed black-box attack). Can only a single network with image transformations and retraining work? In essence we test a single network, with one set of transformations (Guo) and a single network retrained on multiple random transformations (BaRT). Both of these defenses perform worse than BUZz for the mixed black-box attack.

3.3 Buffer Zone Graphs

In Figure 2 we show buffer zone graphs Liu et al. (2017) for different BUZz configurations for a single image in the CIFAR-10 dataset. Each point on the graph represents whether the image is correctly classified (green) wrongly classified (red or blue) or has a null (adversarial) class label. The origin on the graph is the classified image without any noise added. Moving along the x-axis in the graph represents adding more adversarial noise to the image (moving along the positive direction of the gradient). Moving along the y-axis represents adding more random noise to the image. Note creating these graphs can only be done when we have access to the model gradients (i.e. a white-box access) so this is not a way to attack BUZz. The graphs and the technique are merely empirical evidence to show the existence of buffer zones in the BUZz defense. Complete information for generating the graphs are given in the supplemental material.

It is important to note that it may be possible to further combine other defense techniques such as adversarial training, randomizing some of the image transformations or any number of other techniques. However, the goal of this paper is not to exhaustively test every possible defense combination. The goal is not to test every defense in the literature either. The objective of this work is to provide a defense framework against black-box adversaries that offers clear trade-offs between clean accuracy and security.

4 DEFENSE PERFORMANCE METRIC

We introduce a new metric to properly understand the combined effect of both the drop γ in clean accuracy when implementing a defense, as well as the attacker's success rate α against the defense.

1. A drop γ in clean accuracy from an original clean accuracy p to clean accuracy
$$p_d = p - \gamma \tag{1}$$
for the defense. Here, clean accuracy p corresponds to a vanilla scheme without defense strategy
2. The attacker's success rate α against the defense. For an untargeted attack the attacker fails if the defense recognizes an adversarial manipulated image as an adversarial example and outputs the adversarial label $\perp$ or if the defense produces the original class label.

The probability of proper/accurate classification by the defense in the presence of adversaries is equal to $(p-\gamma)(1-\alpha)$ (since the defense properly labels a fraction $p-\gamma$ if no adversary is present and out of these images a fraction α is successfully attacked if an adversary is present). In other words $(p-\gamma)(1-\alpha)$ is the accuracy of the defense in the presence of adversaries (malicious environment). Going from a non-malicious environment without defense to a malicious environment with defense gives a drop in accuracy of

$$\delta = p - (p-\gamma)(1-\alpha) = \gamma + (p-\gamma)\alpha. \tag{2}$$

δ can be used to measure the effectiveness of different defenses, the smaller the better. If two defenses offer roughly the same δ, then it makes sense to consider their (γ, α) pairs and choose the defense that either has the smaller attacker's success rate α or the smaller drop γ in clean accuracy.

From a pure machine learning perspective, in order for a defense to perform well in a non-malicious environment, we want γ very small or, equivalently, p_d close to p. From a pure security perspective, in order for a defense to perform well in a malicious environment, we want δ to be small. Therefore, for properly comparing defenses we focus on tuples $(\delta = \gamma + (p-\gamma)\alpha, p_d = p - \gamma)$, where α corresponds to the best attacker's success rate across the best known attacks from literature.

5 EXPERIMENTAL RESULTS

In this section we provide experimental results to show the effectiveness of the BUZz defense. We experiment with two popular datasets, Fashion-MNIST (Xiao et al., 2017) and CIFAR-10 (Krizhevsky et al.). In terms of network architecture, we use ResNet56 (He et al., 2016) for the networks in the CIFAR-10 defenses and VGG16 (Simonyan & Zisserman, 2014) for the networks in the Fashion-MNIST defenses. Full defense details as well as their standard training procedure can be found in the supplemental material. Unlike other reported results in the literature, for every defense, we construct it using the same network architecture whenever possible, we apply the defense to the same dataset and we run every defense under the same set of attacks. This allows us to provide an unprecedented comparison of adaptive black-box attack results from the literature.

5.1 DEFENSES

We experiment with 11 defenses (BUZz, vanilla thresholding, Guo, Liu, ADP, Xie, Madry, Tramer, Mul-Def, BaRT and Odds). Due to the limited space we cannot describe the full implementation details of every defense here. We encourage the reader to examine the supplemental material for further details if interested.

BUZz and Thresholding Defenses. In this paper we experiment with BUZz and also a naive defense which we call vanilla thresholding. A common misconception is that by merely threholding the output of a vanilla classifier (i.e. marking a sample as adversarial if the network is not confident in its prediction) then all black-box attacks can be mitigated. We provide results for the 70%, 90% and 95% thresholding network to show this is simply not the case.

For BUZz, we realize the buffer zones through image transformations. Specifically, each network has an image transformation selected from mappings $c(x) = Ax + b$. We explain how we chose the

randomized A and b based on the dataset in the supplementary material. We can think of an image transformation $c_j(x)$ as an extra randomly fixed layer added to the layers which form the j-th CNN. We tested three of these designs: One with 8 networks each using a different image resizing operation from 32 to 32, 40, 48, 64, 72, 80, 96, 104. The second with 4 networks being the subset of the 8 networks that use image resizing operations from 32 to 32, 48, 72, 96. The third with 2 networks being a subset of the 8 networks that use image resizing operations from 32 to 32 and 104.

5.2 ATTACKS

On every defense we run pure black-box attacks, mixed black-box attacks and boundary attacks. Each of these attacks can be further categorized based on how the adversarial samples are generated. For the mixed black-box attack (proposed in this paper) we use seven different adversarial generation methods (FGSM, IFGSM, PGD, MIM, C&W and EAD). For pure black-box attacks we use the same set of generations methods (but the model used in conjunction with the attack is not adaptively trained). For the boundary attacks, we consider HSJA and RayS. For CIFAR-10, the maximum perturbation we allow is $\epsilon = 0.05$ and for Fashion-MNIST the maximum perturbation is $\epsilon = 0.1$. For RayS and HSJA we allow 10,000 queries per sample. Note in Table 1 some attacks are not applicable to certain defenses. This occurs only for boundary attacks for 2 defenses (BaRT and Odds). This is due to computational complexity issues of non-parallelizable prediction for the run time of the boundary attacks. We fully explain this in the supplementary material along with precise attack details for all the attacks.

5.3 EXPERIMENTAL ANALYSIS

The main results for our paper are given in Table 1. In this table we compute the δ metric for every defense based on the attack that the defense is weakest to (i.e. has the lowest robust accuracy). For example, if the BUZz-8 defense has a robust accuracy of 60% against RayS (60% of the adversarial samples do not fool the defense) and a robust accuracy of 39% against HSJA, then HSJA is used to compute the BUZz-8 boundary δ metric. Visually the results for the worst case δ metric for the pure and mixed black-box adversaries is given in Figure 3.

In terms of performance, our proposed defense (BUZz) outperforms every other defense for 2 out of the 3 types of attacks, on both CIFAR-10 and Fashion-MNIST. On CIFAR-10, BUZz-4 gives the best trade off between security and accuracy for δ mixed and δ pure. Likewise BUZz-8 does best for these two attacks for Fashion-MNIST. For the boundary attacks, we can see BUZz still does well but both adversarial training (Madry) and randomization based defenses (e.g. Xie) do better. We have not yet experimented with adversarial training with BUZz, or randomization of the BUZz's output. This still leaves many possibilities to make BUZz stronger to boundary attacks in future work. Here we only present BUZz as a stand alone concept to show it can offer security against all attacks, and can outperform all defenses over the majority of black-box attacks.

It is also worth noting we are the first to analyze all of these 10 defenses under this wide spectrum of black-box adversaries. In the supplementary material we go in depth, further analyzing the strengths and weakness of the other 10 defenses.

6 CONCLUSION

In this paper, we advance the field of adversarial machine learning by developing a new metric for measuring defenses (the δ metric), a new attack for testing defenses (the mixed black-box attack), a new defense framework (BUZz), and rigorous experimentation on 11 defenses using 2 datasets, and 10 different black-box attacks. Our new metric helps better understand the cost incurred by the defense, versus the amount of security the defense provides. Our new mixed black-box attack is more effective on certain defenses (like Mul-Def and Xie) than either pure black-box or boundary attacks. Lastly and most importantly, our buffer zone defense offers unprecedented flexibility between security and clean accuracy. It also outperforms other defenses under both mixed black-box and pure black-box attacks.

Table 1: δ values and clean accuracies for all the defenses under different attacks. The best δ for every category is shown in bold. Note robust accuracy for every type of attack (e.g. HSJA, RayS, mixed black-box MIM, pure black-box PGD etc. are given in the supplemental material.

	CIFAR-10				Fashion-MNIST			
	δ Mixed	δ Pure	δ Boundary	Clean Acc	δ Mixed	δ Pure	δ Boundary	Clean Acc
Vanilla	0.6874998	0.5715248	0.9278	0.9278	0.8317484	0.6072044	0.9356	0.9356
Guo	0.523206	0.5222968	0.9278	0.9092	0.5963352	0.385197	0.9356	0.9023
BUZz-2	0.2880736	0.3331607	0.9278	0.8507	0.3414248	0.2603233	0.85023	0.8537
BUZz-4	**0.2120402**	**0.250949**	0.798104	0.8106	0.2653332	0.2513864	0.500788	0.8204
BUZz-8	0.227281	0.253002	0.632765	0.7565	**0.2308226**	**0.2362679**	0.243269	0.7779
Liu	0.3922416	0.3743328	0.9278	0.8528	0.437554	0.33327	0.9356	0.899
ADP	0.789179	0.564745	0.9278	0.943	0.854969	0.6310994	0.9356	0.9486
VanillaT-0.7	0.4551126	0.444267	0.9278	0.9038	0.7768096	0.5672432	0.9356	0.9232
VanillaT-0.95	0.2732236	0.3197976	0.9278	0.8468	0.664432	0.48514	0.9356	0.892
VanillaT-0.99	0.2249932	0.2730551	0.9278	0.7879	0.5126558	0.4113518	0.9356	0.8442
Xie	0.7427232	0.6791472	0.334424	0.7064	0.780484	0.6204696	0.421268	0.8164
Madry	0.3507092	0.5260184	0.536552	0.7524	0.5417105	0.4079975	**0.16232**	0.8055
Tramer	0.5510392	0.5518916	0.9278	0.8524	0.6135816	0.4319782	0.9356	0.9361
MulDef-4	0.6029543	0.614276	0.31817	0.8709	0.60709	0.4503438	0.316124	0.9386
MulDef-8	0.5881268	0.5992496	**0.294656**	0.8556	0.5825395	0.4720325	0.28005	0.9365
BaRT-1	0.5866742	0.5223917	NA	0.8571	0.6291779	0.4944968	NA	0.9039
BaRT-4	0.5949741	0.5934715	NA	0.7513	0.6171104	0.5209	NA	0.8294
BaRT-7	0.6673436	0.698525	NA	0.6114	0.646371	0.5689827	NA	0.7817
BaRT-10	0.7418042	0.7500815	NA	0.4869	0.674844	0.6119768	NA	0.7144
Odds	0.5150502	0.4872003	NA	0.7141	0.8337155	0.6593798	NA	0.7547

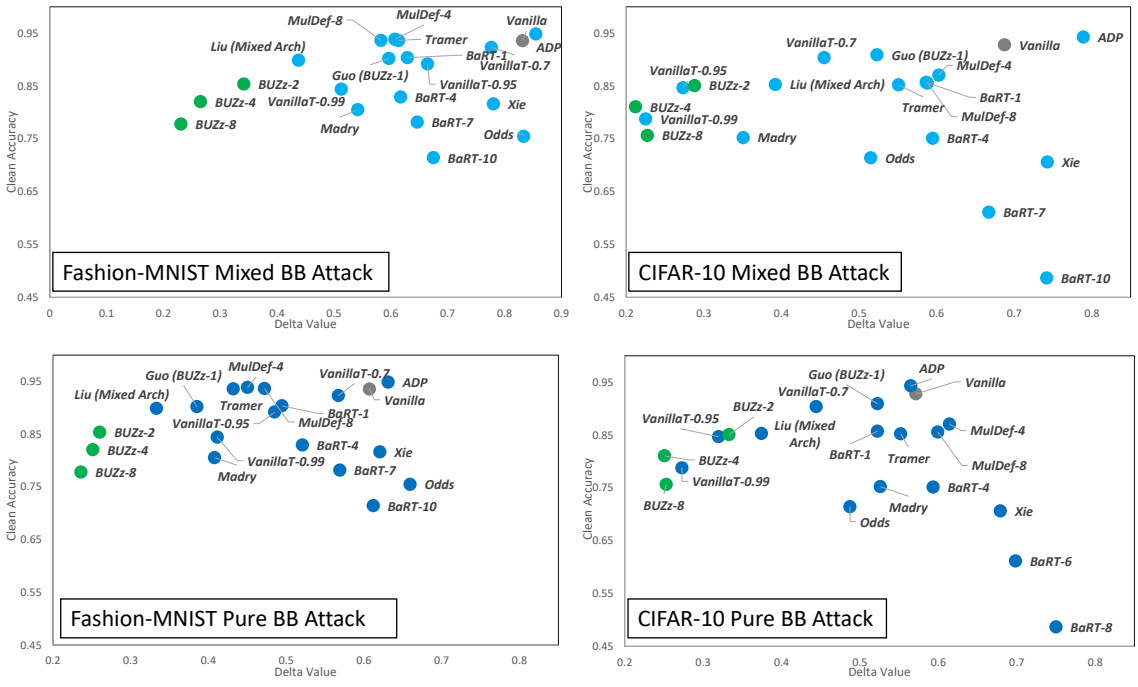

Figure 3: The δ metric vs clean accuracy for the mixed black-box and pure black-box attacks. The BuZz results are shown in green and the vanilla result is shown in gray. Graphs for the boundary attacks are given in the supplemental material.

REFERENCES

Anish Athalye, Nicholas Carlini, and David A. Wagner. Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples. In *ICML 2018*, pp. 274–283, 2018.

Xiaoyu Cao and Neil Zhenqiang Gong. Mitigating Evasion Attacks to Deep Neural Networks via Region-based Classification. *CoRR*, abs/1709.05583, 2017. URL http://arxiv.org/abs/1709.05583.

Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *AISec@CCS*, 2017.

Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019.

Jianbo Chen and Michael I. Jordan. Boundary attack++: Query-efficient decision-based adversarial attack. *CoRR*, abs/1904.02144, 2019. URL http://arxiv.org/abs/1904.02144.

Jianbo Chen, Michael I Jordan, and Martin J Wainwright. Hopskipjumpattack: A query-efficient decision-based attack. In *IEEE S&P*, pp. 1277–1294, 2020.

Jinghui Chen and Quanquan Gu. Rays: A ray searching method for hard-label adversarial attack, 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 *AISec*. ACM, 2017.

Pin-Yu Chen, Yash Sharma, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh. Ead: elastic-net attacks to deep neural networks via adversarial examples. In *Thirty-second AAAI conference on artificial intelligence*, 2018.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 9185–9193, 2018.

Reuben Feinman, Ryan R. Curtin, Saurabh Shintre, and Andrew B. Gardner. Detecting Adversarial Samples from Artifacts. *CoRR*, abs/1703.00410, 2017.

Ross B. Girshick. Fast R-CNN. In *IEEE-ICCV*, pp. 1440–1448, 2015.

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.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *International Con- ference on Learning Representations (ICLR)*, 2015.

Chuan Guo, Mayank Rana, Moustapha Cissé, and Laurens van der Maaten. Countering Adversarial Images using Input Transformations. *CoRR*, 2017.

Chuan Guo, Jacob R Gardner, Yurong You, Andrew Gordon Wilson, and Kilian Q Weinberger. Simple black-box adversarial attacks. *arXiv preprint arXiv:1905.07121*, 2019.

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.

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). URL http://www.cs.toronto.edu/~kriz/cifar.html.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (eds.), *NIPS*, pp. 1097–1105. 2012.

Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial Machine Learning at Scale. *CoRR*, abs/1611.01236, 2016. URL http://arxiv.org/abs/1611.01236.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *International Conference on Learning Representations (ICLR) Workshop*, 2017.

Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into Transferable Adversarial Examples and Black-box Attacks. *ICLR (Poster)*, 2017.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations (ICLR)*, 2018.

Dongyu Meng and Hao Chen. Magnet: a two-pronged defense against adversarial examples. In *CSS*, pp. 135–147. ACM, 2017.

Jan Hendrik Metzen, Tim Genewein, Volker Fischer, and Bastian Bischoff. On Detecting Adversarial Perturbations. *CoRR*, abs/1702.04267, 2017. URL http://arxiv.org/abs/1702.04267.

Tianyu Pang, Kun Xu, Chao Du, Ning Chen, and Jun Zhu. Improving Adversarial Robustness via Promoting Ensemble Diversity. In *ICML*, pp. 4970–4979, 2019.

Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE Symposium on Security and Privacy (SP)*, pp. 582–597. IEEE, 2016a.

Nicolas Papernot, Patrick D. McDaniel, and Ian J. Goodfellow. Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples. *CoRR*, 2016b.

Nicolas Papernot, Patrick D. McDaniel, Ian J. Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical Black-Box Attacks against Machine Learning. In *ACM AsiaCCS 2017*, pp. 506–519, 2017.

Edward Raff, Jared Sylvester, Steven Forsyth, and Mark McLean. Barrage of random transforms for adversarially robust defense. In *CVPR*, pp. 6528–6537, 2019.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NIPS*, pp. 91–99, 2015.

Kevin Roth, Yannic Kilcher, and Thomas Hofmann. The odds are odd: A statistical test for detecting adversarial examples. 2019.

Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(4):640–651, April 2017. ISSN 0162-8828.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *ICLR*, 2015.

Siwakorn Srisakaokul, Zexuan Zhong, Yuhao Zhang, Wei Yang, and Tao Xie. Muldef: Multi-model-based defense against adversarial examples for neural networks. *arXiv preprint arXiv:1809.00065*, 2018.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *CoRR*, abs/1312.6199, 2013. URL http://arxiv.org/abs/1312.6199.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *ICLR*, 2014.

Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Dan Boneh, and Patrick D. McDaniel. Ensemble Adversarial Training: Attacks and Defenses. *CoRR*, abs/1705.07204, 2017. URL http://arxiv.org/abs/1705.07204.

Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. *arXiv preprint arXiv:2002.08347*, 2020.

Jianyu Wang, Zhishuai Zhang, Cihang Xie, Yuyin Zhou, Vittal Premachandran, Jun Zhu, Lingxi Xie, and Alan L. Yuille. Visual Concepts and Compositional Voting. *CoRR*, 2017.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. 2017. URL http://arxiv.org/abs/1708.07747.

Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. *ICLR (Poster)*, 2018.

Xiaoyong Yuan, Pan He, Qile Zhu, Rajendra Rana Bhat, and Xiaolin Li. Adversarial Examples: Attacks and Defenses for Deep Learning. *CoRR*, abs/1712.07107, 2017. URL http://arxiv.org/abs/1712.07107.