

# Model Agnostic Dual Quality Assessment for Adversarial Machine Learning and an Analysis of Current Neural Networks and Defenses

Danilo Vasconcellos Vargas, Shashank Kotyan

**Abstract**—In adversarial machine learning, there are a huge number of attacks of various types which makes the evaluation of robustness for new models and defenses a daunting task. To make matters worse, there is an inherent bias in attacks and defenses. Here, we organize the problems faced (model dependence, insufficient evaluation, unreliable adversarial samples and perturbation dependent results) and propose a dual quality assessment method together with the concept of robustness levels to tackle them. We validate the dual quality assessment on state-of-the-art models (WideResNet, ResNet, AliConv, DenseNet, NIN, LeNet and CapsNet) as well as the current hardest defenses proposed at ICLR 2018 as well as the widely known adversarial training, showing that current models and defenses are vulnerable in all levels of robustness. Moreover, we show that robustness to  $L_0$  and  $L_\infty$  attacks differ greatly and therefore duality should be taken into account for a correct assessment. Interestingly, a by-product of the assessment proposed is a novel  $L_\infty$  black-box method which requires even less perturbation than the One-Pixel Attack (only 12% of One-Pixel Attack's amount of perturbation) to achieve similar results. Thus, this paper elucidates the problems of robustness evaluation, proposes a dual quality assessment to tackle them as well as analyze the robustness of current models and defenses. Hopefully, the current analysis and proposed methods would aid the development of more robust deep neural networks and hybrids alike.

Code available at: <http://bit.ly/DualQualityAssessment>

**Index Terms**—Adversarial Machine Learning, Robustness Evaluation, One-Pixel Attack, Deep Neural Networks, ResNet, CapsNet, Dual Quality Assessment

## I. INTRODUCTION

Deep Neural Networks (DNN) has allowed us to achieve high accuracy in speech recognition, face recognition, among other applications. In fact, most of these applications are only possible through the use of DNNs. Despite these achievements, DNNs were shown to misclassify when small perturbations are added to original samples, called adversarial samples.

In fact, security as well as safety risks are prohibiting the use of machine learning, specially DNNs, in many important applications such as autonomous vehicles. Therefore, it is of utmost importance to create not only accurate but robust machine learning. However, to do so a quality assessment is needed which would allow robustness to be checked easily without a deep knowledge of adversarial machine learning.

Moreover, adversarial samples point out to reasoning shortcomings in machine learning. In other words, it reveals that current methods are not able to understand concepts or high-level abstractions as we once thought. Improvements in robustness should also result in learning systems that can better reason over data as well as achieve a new level of abstraction. Therefore, a quality assessment procedure would also be helpful in this regard, checking for failures in both reasoning and high-level abstractions.

Regarding the development of a quality assessment for robustness, adversarial machine learning has provided some tools which could be useful for the development. However, the sheer amount of scenarios, attacking methods and metrics ( $L_0$ ,  $L_1$ ,  $L_2$  and  $L_\infty$ ) make the current state-of-the-art difficult to grasp. Giving the huge amount of possibilities and many definitions with their exceptions and trade-offs, it turns what should be a simple robustness quality assessment into a daunting task. To make matters worse, most of the current attacks are white box ones which cannot be used to evaluate hybrids, non-standard DNNs and other classifiers in general. Therefore, to create a quality assessment procedure there are a number of problems that must be tackled:

- **P1 - Model Dependence** - To allow DNNs to be compared with other approaches which may be completely different from current DNNs (logic hybrids, evolutionary hybrids and other learning systems) a model agnostic quality assessment is necessary.
- **P2 - Insufficient Evaluation** - There are many types of adversarial samples as well as possible attack variations and scenarios each with their own bias. The attacks also differ substantially depending on metrics optimized, namely  $L_0$ ,  $L_1$ ,  $L_2$  and  $L_\infty$ . However, not all of them are essential for the evaluation of robustness. A quality assessment should have few but sufficient tests to allow for a deep analysis without compromising its use.
- **P3 - Unreliable Adversarial Samples** - Some attacks are known to produce unrecognizable adversarial samples under certain scenarios which can only be discovered through inspection. Both the need for inspection together with the possibility of unrecognizable adversarial samples should not be present.
- **P4 - Perturbation Dependent Results** - Different amount of perturbation leads to different attack accuracy. Moreover, models are more or less susceptible to different amount of perturbation and might surpass some models

D. V. Vargas is with the Faculty of Information Science and Electrical Engineering, Kyushu University, Fukuoka, Japan (email: [vargas@inf.kyushu-u.ac.jp](mailto:vargas@inf.kyushu-u.ac.jp))

S. Kotyan is with the Department of Computer Science and Engineering, Dr. SPM IIIT-NR, India (email: [shashank15100@iiitnr.edu.in](mailto:shashank15100@iiitnr.edu.in))

for some levels of perturbation while being surpassed otherwise. Consequently, this might result in double standards or hide important information.

In this paper, we propose a quality assessment which tackles all of the problems above. It has the following features:

- 1) **Non-gradient based Black-box Attack (Address P1)** - To allow for model agnostic evaluation which do not depend on specific features of the learning process such as gradients, black-box attacks are desirable. Therefore, here the proposed quality assessment is based on black-box attacks, one of which is a novel  $L_\infty$  black-box attack. In fact, to the knowledge of the authors, this is the first  $L_\infty$  black box attack that does not make any assumptions over the target machine learning system. Both of the attacks used are based on evolutionary algorithms which are known to achieve state-of-the-art results in black-box optimization. Figures 1 and 2 show some adversarial samples crafted.
- 2) **Dual Evaluation (Address P2 and P3)** - Here we focus on black-box attacks as well as use a dual evaluation approach with a method based on  $L_0$  and another one based on  $L_\infty$ . In this manner, the evaluation still contain the two attack extremes preserving different attack vectors without adding much overhead to the evaluation, i.e., the evaluation use attacks that either perturb a few pixels strongly ( $L_0$ ) or all pixels slightly ( $L_\infty$ ). This choice also eliminates a lot of problems from other metrics discussed in Section III, such as the chance producing unrecognizable adversarial samples (Problem P3).
- 3) **Robustness Levels (Address P4)** - In this paper, we define robustness levels in terms of the constraint's threshold  $th$  (Equations 1 and 2) and compare multiple levels of results with their respective values in the same level. This avoids the comparison of results with different degrees of perturbation (Problem P4) and allow for robustness to be evaluated not as a global value but in relative degrees. In fact, robustness levels add a concept which may aid in the classification of algorithms, e.g., an algorithm which is robust to one pixel attack belongs to the 1-pixel-safe category.

## II. RELATED WORK

Adversarial machine learning can be seen as a constrained optimization problem. Before defining it, let us formalize adversarial samples first. Let  $f(\mathbf{x}) \in \mathbb{R}^k$  be the output of a machine learning algorithm denoted by function  $f$  in which  $\mathbf{x} \in \mathbb{R}^{m \times n \times 3}$  is the input of the algorithm for input and output of respective sizes  $m \times n \times 3$  (images with three channels are considered) and  $k$ . Adversarial samples  $\mathbf{x}'$  are explicitly defined as follows:

$$\begin{aligned} \mathbf{x}' &= \mathbf{x} + \epsilon_{\mathbf{x}} \\ \{\mathbf{x}' \in \mathbb{R}^{m \times n \times 3} \mid \underset{j}{\operatorname{argmax}}(f(\mathbf{x}')) &\neq \underset{i}{\operatorname{argmax}}(f(\mathbf{x}))\}, \end{aligned} \quad (1)$$

where  $\epsilon_{\mathbf{x}} \in \mathbb{R}^{m \times n \times 3}$  is a small perturbation added to the input. Making use of the definition of adversarial samples,

adversarial machine learning can be defined as the following optimization problem for untargeted black-box attacks:

$$\begin{aligned} \underset{\epsilon_{\mathbf{x}}}{\text{minimize}} \quad & f(\mathbf{x} + \epsilon_{\mathbf{x}})_c \\ \text{subject to} \quad & \|\epsilon_{\mathbf{x}}\| \leq th \end{aligned} \quad (2)$$

in which  $f(\cdot)_c$  denotes the soft label for the correct class  $c$  while  $th$  is a threshold value.

The constraint in the optimization problem has the objective of disallowing perturbations which could make  $\mathbf{x}$  unrecognizable or change its correct class. Therefore, the constraint is itself a mathematical definition of what constitutes an imperceptible perturbation. Many different norms are used in the literature (e.g.,  $L_0$ ,  $L_1$ ,  $L_2$  and  $L_\infty$ ). Intuitively, the norms allow for different types of attacks.  $L_0$  allows attacks to perturb a few pixels strongly,  $L_\infty$  allow all pixels to change slightly and both  $L_1$  and  $L_2$  allow for a mix of both strategies.

### A. Recent Advances in Attacks and Defenses

Recently, the DNNs were shown to share many vulnerabilities. The first paper on the subject dates back to 2013 when DNNs were shown to behave strangely for nearly the same images [1]. Afterwards, a series of vulnerabilities were found. In [2], the authors demonstrated that DNNs show high confidency to textures and random noise. Single adversarial perturbations which can be added to most of the samples to fool a DNN was shown to be possible [3]. Patches can also make them misclassify and the addition of them in an image turn it into a different class [4]. Moreover, an extreme attack was shown to be effective. It was shown that it is possible to make DNNs misclassify with a single pixel change [5].

In fact, many of these attacks can be easily made into real world threats as shown in [6], i.e., printed out adversarial samples still work because many adversarial samples are robust against different light conditions. Moreover, carefully crafted glasses can also be made into attacks [7] or even general 3d adversarial objects were shown possible [8].

Many defensive systems and detection systems were proposed to mitigate some of the problems. However, there are still no current solutions or promising ones. Regarding defensive systems, defensive distillation in which a smaller neural network squeezes the content learned by the original one was proposed as a defense [9] however it was shown to not be robust enough in [10]. Adversarial training was also proposed in which adversarial samples are used to augment the training dataset in such a way that the DNN will be able to correctly classify them, increasing its robustness [11], [12], [13]. Although adversarial training can increase slightly the robustness it is still vulnerable to attacks [14]. There are many recent variations of defenses [15], [16] which are carefully analyzed and many of their shortcomings are explained in [17], [18].

Regarding detection systems, a study from [19] demonstrated that indeed some adversarial samples have different statistical properties which could be exploited for detection. In [20], the authors propose to compare the prediction of a classifier with a prediction of the same input but "squeezed"

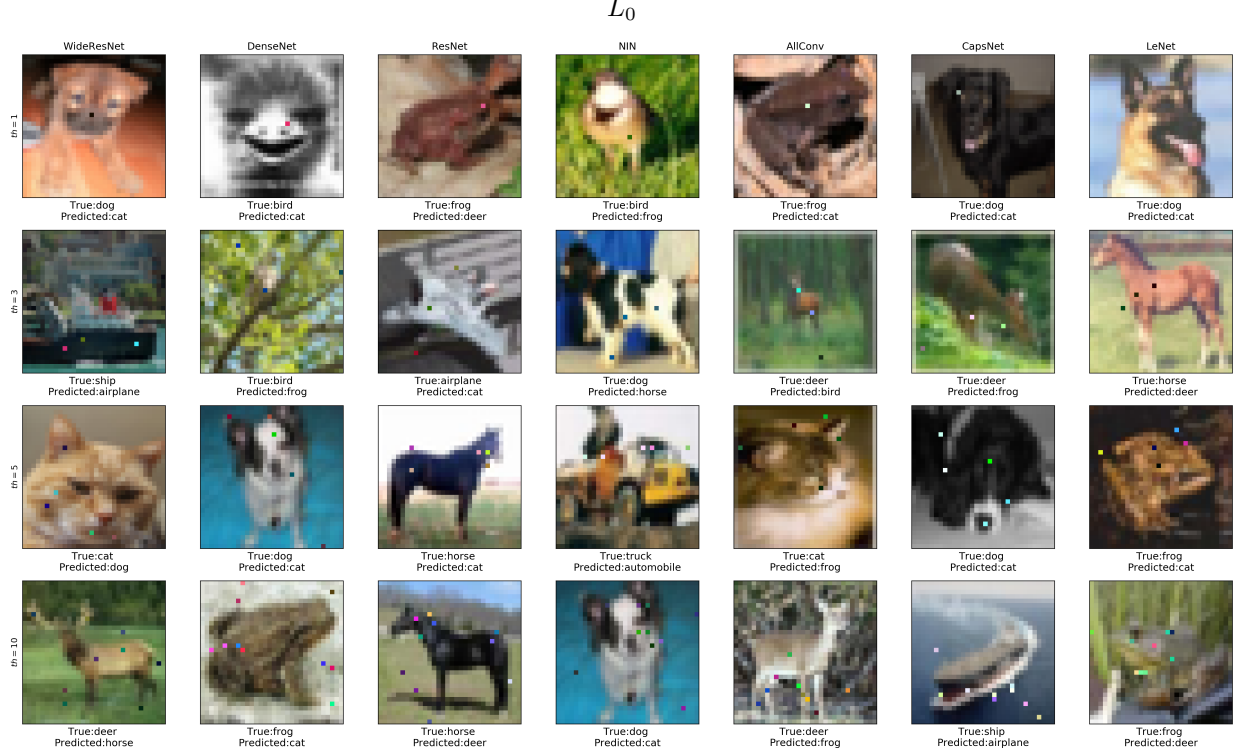


Fig. 1. Adversarial samples found with few-pixel black-box attack ( $L_0$ ) for state-of-the-art DNNs on CIFAR.

(either color or spatial smoothing). This allow classifiers to detect adversarial samples with small perturbations. Having said that, many detection systems might fail when adversarial samples differ from test conditions [21], [22]. Thus, the clear benefits of detection systems remains inconclusive.

Few works focus on a less direct objective. The one of not attacking or defending but understanding the reason behind such lack of robustness. In [11] it is argued that DNNs' linearity are one of the main reasons. Another recent investigation proposes a technique for evaluating the influences of small perturbations called propagation maps as well as propose the conflicting saliency hypothesis as a reason for adversarial samples' existence [23]. The hypothesis is based on a conflict between saliency from the original natural image and the saliency inserted by adding an adversarial perturbation.

### III. DUAL QUALITY ASSESSMENT

In this paper, we propose a dual quality assessment procedure to evaluate the robustness of general machine learning algorithms. By using both  $L_0$  and  $L_\infty$  black-box attacks the proposed quality assessment is able to measure the robustness accurately for any algorithm. They are described in detail below.

The reasoning behind the choice of using  $L_0$  and  $L_\infty$  are as follows. Without altering much the original sample, attacks can perturb a few pixels strongly ( $L_0$ ), all pixels slightly ( $L_\infty$ ) or a mix of both ( $L_1$  and  $L_2$ ). The hurdle is that  $L_1$  and  $L_2$  which mix both strategies vary strongly with the size of images, if not used with caution may cause unrecognizable adversarial samples (Problem P3) and is also difficult to compare between

methods because the amount of perturbations will often differ (Problem P4). Moreover,  $L_1$  and  $L_2$  are only mixing the other metrics and does not imply in a new attack vector. Therefore we focus on using both  $L_0$  and  $L_\infty$  as the dual attack vectors.

#### A. Threshold Attack ( $L_\infty$ black-box attack)

The threshold attack optimizes the constrained optimization problem defined in Equation 2 with the constraint  $\|\epsilon_x\|_\infty \leq th$ , i.e., it uses the  $L_\infty$  norm.  $th$  is a threshold which is set here to be one of the following values  $\{1, 3, 5, 10\}$ .

The search space is exactly the same as the input space because the variables can be any variation of the input as long as the threshold is respected. Therefore, the algorithm search in  $\mathbb{R}^{m \times n \times 3}$  space.

The threshold attack proposed here uses the state-of-the-art black-box optimization algorithm called covariance matrix adaptation evolution strategy (CMA-ES) [24]. Here we use the canonical version of the algorithm to have a clear standard. To satisfy the constraint a simple repair method is employed in which pixels that surpass the minimum/maximum are brought back to the minimum/maximum value. Notice that the optimization uses real values and that pixel values are kept within range with a clipping function.

#### B. Few-Pixel Attack ( $L_0$ black-box attack)

The few-pixel attack optimizes the constrained optimization problem defined in Equation 2. However, the constraint used is  $\|\epsilon_x\|_0 \leq th$ , i.e., it uses the  $L_0$  norm.  $th$  is a threshold which may also admit one of the following values  $\{1, 3, 5, 10\}$ .



Fig. 2. Adversarial samples found with threshold black-box attack ( $L_\infty$ ) for state-of-the-art DNNs on CIFAR.

The search variable is a combination of pixel values (3 values) and position (2 values) for all of the pixels ( $th$  pixels). Therefore, the search space is smaller than the threshold attack with dimensions of  $\mathbb{R}^{5 \times th}$ . To conduct the optimization we use the CMA-ES which is a widely known black-box optimization algorithm. The constraint is always satisfied because the number of parameters is itself modeled after the constraint. In other words, when searching for one pixel perturbation, the number of variables are fixed to pixel values (three values) plus position values (two values). Therefore it will always modify only one pixel, respecting the constraint. Since the optimization is done in real values, to force the values to be within range a simple clipping function is used for pixel values. For position values a modulo operation is executed.

#### IV. ROBUSTNESS LEVELS

Machine learning algorithms might behave differently to varied amount of perturbations. To cope with the relativeness of robustness, here we propose robustness levels. Robustness levels evaluate classifiers in a couple of  $th$  thresholds. Specifically, we define four levels of robustness 1, 3, 5, 10 for both  $L_0$  and  $L_\infty$  and name them respectively pixel and threshold robustness levels. Algorithms that pass a level of robustness (0% attack accuracy) are called level-threshold-safe or level-pixel-safe, substituting the word level by the level it passes. For example, an algorithm that passes the level one in threshold ( $L_\infty$ ) attacks is called 1-threshold-safe.

#### V. EXPERIMENTS

In this section, we aim to validate the dual quality assessment empirically as well as analyze the current state-of-the-art in robustness. Therefore, the following tests are conducted:

- **Preliminary Tests** - Tests on two state-of-the-art DNNs are presented (ResNet [25] and CapsNet [26]). These tests are done to choose the black-box optimization algorithm to be used for the next sections. The performance of both Differential Evolution (DE) [27] and CMA-ES are evaluated.
- **Evaluating Learning Systems** - Tests are extended to seven different DNNs of the state-of-the-art - WideResNet [28], DenseNet [29], ResNet [25], Network in Network (NIN) [30], All Convolutional Network (AllConv) [31], CapsNet [26] and LeNet [32]. In this section, the total attack accuracy as well as the attack accuracy per class are evaluated to investigate possible reasons for robustness.
- **Evaluating Defense Systems** - Attack accuracy on two state-of-the-art defenses published on ICLR 2018 as well as the widely known adversarial training defense are conducted and analyzed. We have chosen defenses based on completely different principles to be tested. In this way, the results achieved here can be extended to other similar types of defenses in the literature.
- **Extremely Fast Quality Assessment** - In this section we verify the possibility of an extremely fast version of the proposed quality assessment. Instead of a full-fledged optimization, already crafted adversarial samples are used to fool other models and defenses. This would

enable attacks to have a  $O(1)$  time complexity, being significantly faster and is similar to the transferability of adversarial samples.

- **Quality Assessment's Attack Distribution** - To further evaluate the dual attack distribution as well as demonstrate the necessity of such duality, the distribution of successful attacks are shown and previous attacks are analyzed in this perspective.
- **Accuracy per Threshold Curve** - This section aims to analyze the complete behavior of the attack accuracy per threshold without restricting the  $th$ 's value. In this manner, we can also verify if results of the previous sections using fixed  $th$  are a good approximation to the curve.

### A. Settings

Pixel values that exceed 255 are clipped to remain in valid range. Search variables related to position values are also clipped to remain inside the minimum/maximum allowed. The parameters of both CMA-ES and DE are described respectively in Tables I and II. DE's use a repair method in which values that goes beyond range are set to random points within the valid range.

Parameter	$L_0$ Attack	$L_\infty$ Attack
Parameter Size	$th * 5$	3072
Function Evaluations	40000	39200
Sigma	31.75	$th / 4$

TABLE I  
PARAMETERS FOR CMA-ES

Parameter	$L_0$ Attack	$L_\infty$ Attack
Parameter Size	$th * 5$	3072
Population Size	400	3072
Number of Generations	100	100
Crossover Probability	1	1

TABLE II  
PARAMETERS FOR DE

### B. Preliminary Tests - Choosing the Optimization Algorithm

Table III shows the accuracy results. Both black-box attacks are able to craft adversarial samples in all levels of robustness. This demonstrates that without knowing anything about the model or learning system and in a very limited setting (few pixels or small threshold change), black-box attacks are still able to reach more than 80% attack accuracy in state-of-the-art DNNs.

Regarding the comparison of CMA-ES and DE, the results justify the choice of CMA-ES for the quality assessment. Both CMA-ES and DE perform similarly in the few-pixel attack scenario. With both DE and CMA-ES having the same number of wins. In the threshold scenario, however, the performance varies greatly. CMA-ES this time always wins (eight wins) against DE (no win). This domination of CMA-ES is expected since the threshold attack has a high dimensional search space which is more suitable for CMA-ES. This happens in part

		$L_0$ Attack's $th$			
		1	3	5	10
ResNet	(DE)	<b>24%</b>	<b>70%</b>	<b>75%</b>	79%
	(CMA-ES)	12%	52%	73%	<b>85%</b>
CapsNet	(DE)	<b>21%</b>	37%	<b>49%</b>	<b>57%</b>
	(CMA-ES)	20%	<b>39%</b>	40%	41%

		$L_\infty$ Attack's $th$			
		1	3	5	10
ResNet	(DE)	5%	23%	53%	82%
	(CMA-ES)	<b>33%</b>	<b>71%</b>	<b>76%</b>	<b>83%</b>
CapsNet	(DE)	11%	13%	15%	23%
	(CMA-ES)	<b>13%</b>	<b>34%</b>	<b>72%</b>	<b>97%</b>

TABLE III  
ACCURACY RESULTS FOR FEW-PIXEL ATTACK ( $L_0$  BLACK-BOX ATTACK) AND THRESHOLD ATTACK ( $L_\infty$  BLACK-BOX ATTACK). LEFT COLUMN SHOWS THE MODEL ATTACKED WITH THE OPTIMIZATION ALGORITHM USED BETWEEN BRACKETS. THE ATTACK IS PERFORMED OVER 100 CORRECTLY CLASSIFIED SAMPLES.

because DE's operators may allow some variables to converge prematurely. CMA-ES, on the other hand, is always generating slightly different solutions while evolving a distribution.

In these preliminary tests, CapsNet was shown overall superior to ResNet. Few-pixel attacks ( $L_0$  attacks) reach 85% attack accuracy for ResNet when ten pixels are modified. CapsNet, on the other hand, is more robust to few-pixel attacks, allowing them to reach only 52% and 41% attack accuracy when ten pixels are modified for DE and CMA-ES respectively. Having said that, CapsNet is less robust than ResNet to the threshold attack with  $th = 10$  in which almost all images were vulnerable (97%) while being at the same time reasonably robust to 1-threshold-safe (only 13% attack accuracy). ResNet is almost equally not robust throughout, with low robustness even when  $th = 3$ , losing to CapsNet in robustness in all other values of  $th$  of the threshold attack. This preliminary tests also shows that different networks have different robustness not only in regard to the type of attacks ( $L_0$  and  $L_\infty$ ) but also in relation to the degree of attack (e.g., 1-threshold and 10-threshold attacks have very different results on CapsNet).

### C. Evaluating Learning Systems

Table IV extends the CMA-ES attacks on other DNNs - WideResNet [28], DenseNet [29], ResNet [25], Network in Network (NIN) [30], All Convolutional Network (AllConv) [31], CapsNet [26] and LeNet [32]. Figures 1 and 2 show examples of crafted adversarial samples. Here, taking into account an existing variance of results, we consider results within five of the lowest to be equally good. These results are written in bold. If we consider the number of bold results for each of the DNNs, a qualitative measure of robustness. CapsNet and AllConv can be considered the most robust with five bold results. The third place in robustness achieves only three bold results and therefore is far away from the top performers.

Notice that none of the DNNs were able to reduce low  $th$  attacks to zero. This demonstrates that although robustness may differ between current DNNs, none of them are able

Model (Accuracy)	$L_0$ Attack's $th$			
	1	3	5	10
WideResNet (95.12)	<b>11%</b>	55%	75%	94%
DenseNet (94.54)	<b>9%</b>	43%	66%	78%
ResNet (92.67)	<b>12%</b>	52%	73%	85%
NIN (90.87)	18%	62%	81%	90%
AllConv (88.46)	<b>11%</b>	<b>31%</b>	57%	77%
CapsNet (79.03)	21%	37%	<b>49%</b>	<b>57%</b>
LeNet (73.57)	58%	86%	94%	99%

Model (Accuracy)	$L_\infty$ Attack's $th$			
	1	3	5	10
WideResNet (95.12)	15%	97%	98%	100%
DenseNet (94.54)	23%	68%	<b>72%</b>	<b>74%</b>
ResNet (92.67)	33%	71%	<b>76%</b>	83%
NIN (90.87)	<b>11%</b>	86%	88%	92%
AllConv (88.46)	<b>9%</b>	70%	<b>73%</b>	<b>75%</b>
CapsNet (79.03)	<b>13%</b>	<b>34%</b>	<b>72%</b>	97%
LeNet (73.57)	44%	96%	100%	100%

TABLE IV

ATTACK ACCURACY RESULTS FOR FEW-PIXEL ATTACK ( $L_0$  BLACK-BOX ATTACK) AND THRESHOLD ATTACK ( $L_\infty$  BLACK-BOX ATTACK). LEFT COLUMN SHOW THE MODEL ATTACKED WITH THE CLASSIFICATION ACCURACY BETWEEN BRACKETS. THE ATTACK IS PERFORMED OVER 100 RANDOM SAMPLES USING THE CMA-ES OPTIMIZATION ALGORITHM. RESULTS IN BOLD ARE THE LOWEST ATTACK ACCURACY AND OTHER RESULTS WHICH ARE WITHIN A DISTANCE OF FIVE FROM THE LOWEST ONE.

to completely overcome even the lowest level of perturbation possible.

The behavior of  $L_0$  and  $L_\infty$  differs specially in the most robust DNNs. Showing that the robustness is achieved with some trade-offs. Moreover, this further justify the importance of using both metrics to evaluate DNNs.

To evaluate the dependence of attacks on specific classes, on Table 3 we further separated the attack accuracy (Table IV) into classes. This table shows an already known feature, that some classes are easier to attack than others. For example, the columns for bird and cat classes are visually darker than frog and truck classes for all diagrams. This happens because classes with similar features and therefore closer decision boundaries are easier to attack.

Interestingly, a close look at Table 3 reveals that robust DNNs tend to be harder to attack in only a few classes. This may suggest that these DNNs encode some classes far away from others (e.g., project the features of these classes into a very different vector). Consequently, the reason of their relative robustness may lie on a simple construction of the decision boundary with a few distinct and strongly separated classes.

#### D. Evaluating Defense Systems

In this section, three defenses proposed in 2018 are evaluated: adversarial training (AT) [13], total variance minimization (TVM) [16], and feature squeezing (FS) [20]. While adversarial training trains the network on adversarial samples, TVM modify the input image by removing a few pixels and reconstructing the image with small total variation. In the adversarial model trained, the model from [13] was downloaded and used. The model has an original accuracy of 87.11% in the CIFAR test dataset. For TVM, we trained a ResNet on

	$L_0$ Attack's $th$			
	1	3	5	10
AT (87%)	22% (67%)	52% (41%)	66% (29%)	86% (12%)
TVM (47%)	16% (39%)	12% (41%)	20% (37%)	24% (35%)
FS (92%)	17% (72%)	49% (44%)	69% (26%)	78% (19%)

	$L_\infty$ Attack's $th$			
	1	3	5	10
AT (87%)	3% (84%)	12% (76%)	25% (65%)	57% (37%)
TVM (47%)	4% (45%)	4% (45%)	6% (44%)	14% (40%)
FS (92%)	26% (64%)	63% (32%)	66% (29%)	74% (22%)

TABLE V

ACCURACY RESULTS FOR FEW-PIXEL ( $L_0$ ) AND THRESHOLD ATTACK ( $L_\infty$  BLACK-BOX ATTACK) ON ADVERSARIAL TRAINING (AT) [13], TOTAL VARIANCE MINIMIZATION (TVM) [16] AND FEATURE SQUEEZING (FS) [20] DEFENSES. THE ORIGINAL (LEFTMOST COLUMN) AND UNDER ATTACK ACCURACY OF THE DEFENSES ARE BETWEEN BRACKETS. FOR THE MODIFIED ACCURACY THE VALUE IS CALCULATED BY MULTIPLYING THE ORIGINAL ACCURACY BY ONE MINUS ATTACK ACCURACY. THE ATTACK IS PERFORMED OVER 100 CORRECTLY CLASSIFIED SAMPLES.

TVM modified images and, albeit many trials with different hyper-parameters, we were able to craft a classifier with at best 47.55% accuracy. This is a steep drop from the 92.37% accuracy of the original ResNet and happens because TVM was originally conceived for Imagenet and does not scale well to CIFAR. For FS, we trained a ResNet on FS modified images and were able to craft a classifier with 92.37% accuracy.

Table V show the few-pixel and threshold attacks' accuracy on adversarial training [13], total variance minimization (TVM) defenses [16] and feature squeezing (FS) [20]. Regarding the adversarial training, it is easier to attack with the few-pixel attack than with threshold attack. This should derive from the fact that the adversarial samples used in the training contained mostly images from  $L_\infty$  type of attacks. This happens because PGD, which was used to create the adversarial samples used in the adversarial training, is an  $L_\infty$  attack. Therefore, this demonstrates that, currently, *given an attack bias that differ from the invariance bias used to train the networks, the attack can easily succeed*. Regarding TVM, the attacks were less successful but the original accuracy of the model trained with TVM is also not great. Therefore, even with a small attack percentage of 24% the resulting model accuracy is 35%. Attacks on Feature Squeezing had a relatively high accuracy. This is true for both  $L_0$  and  $L_\infty$  attacks. Actually, both types of attacks had similar accuracy, revealing a lack of bias in the defense system.

#### E. Extremely Fast Quality Assessment (Transferability)

If adversarial samples from one model can be used to attack different models and defenses, it would be possible to create an ultra fast quality assessment. Figure 4 shows that indeed it is possible to qualitatively assess a model/defense based on the transferability of adversarial samples.

Beyond being a faster method, the transfer of samples have the benefit of ignoring any masking of gradients which makes hard to search but not to transfer because the vulnerability is still there but hidden. Interestingly, the transferability is mostly independent on the type of attack ( $L_0$  or  $L_\infty$ ), with most of the previous discussed differences disappearing. Having said that,



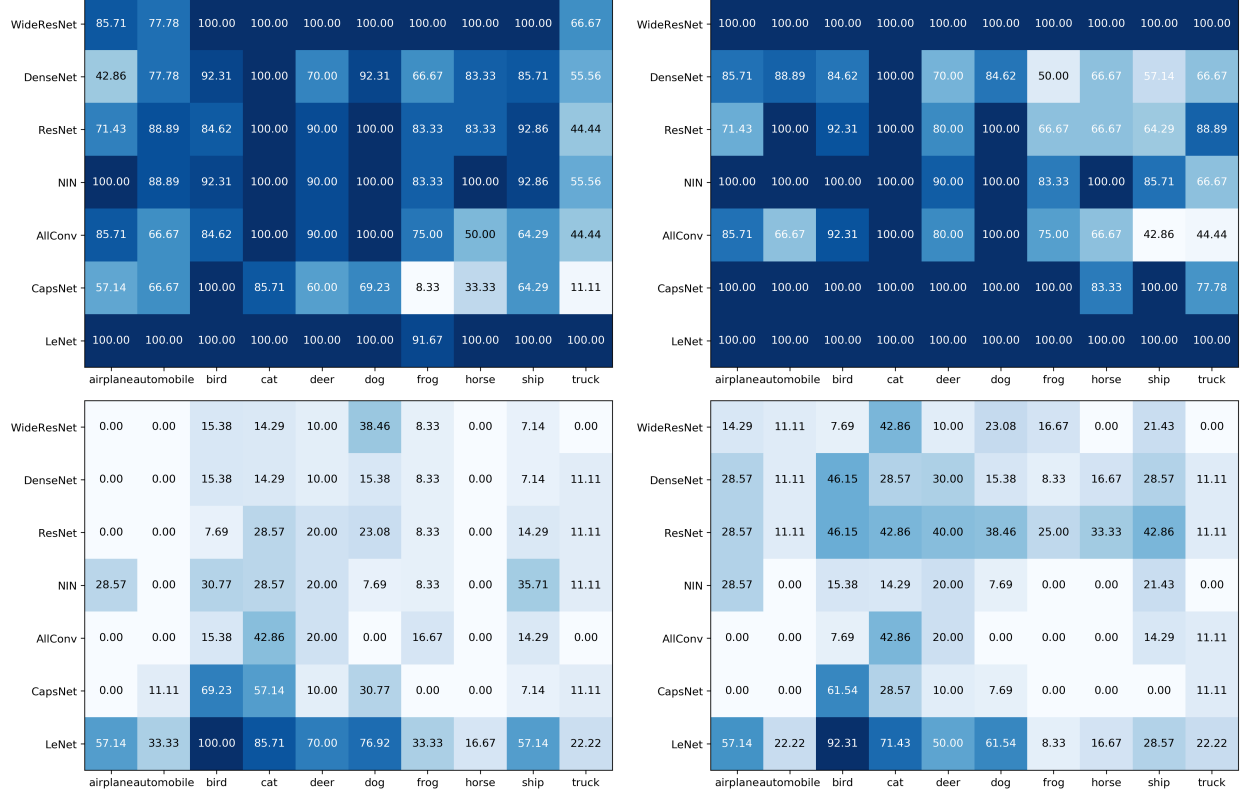


Fig. 3. Attack accuracy from Table IV across classes. The two diagrams at left and right are respectively  $L_0$  and  $L_\infty$  attacks. The top diagrams used  $th = 10$  while the bottom ones used  $th = 1$ .

there are some differences like  $L_0$  attacks are less accurate than most of the  $L_\infty$  ones. This suggests that positions of pixel and their variance are relatively more model specific than small changes in the whole image.

Generally speaking, transferability is a fast assessment method which, when used with many different types of adversarial samples, gives an approximation of the model's robustness. This approximation is not better or worse but different. It differs from usual attacks because (a) it is not affected by how difficult it is to search adversarial samples, taking into account only their existence, and (b) it measures the accuracy to commonly found adversarial samples rather than all searchable ones.

Therefore, in the case of low  $th$  values, transferability can be used as a qualitative measure of robustness. However, its values are not equivalent or close to real attack accuracy. Thus it serves only as a lower bound.

#### F. Quality Assessment's Attack Distribution

This section aims to verify the importance of the duality for the proposed quality assessment by analyzing the distribution of attacks. In some cases, the distribution of samples for  $L_0$  and  $L_\infty$  can be easily verified by the difference in attack accuracy. For example, CapsNet is more susceptible to  $L_\infty$  than  $L_0$  types of attacks (Table III) while for adversarial training [13] the opposite is true (Table V). Naturally, adversarial training depends strongly on the adversarial samples

used in the training and therefore depending on the adversarial samples used a different robustness could be acquired.

Moreover, we show here that even when accuracy seem close the distribution of  $L_0$  and  $L_\infty$  attacks may differ. For example, the attack accuracy on ResNet for both  $L_0$  and  $L_\infty$  with  $th = 10$  differ by mere 2%. However, the distribution of adversarial samples shows that around 17% of the samples can only be attacked by either one of the attack types (Figure 5).

Thus, the evaluation of both  $L_0$  and  $L_\infty$  are important to verify the robustness of a given model or defense and this is true even when a similar accuracy is observed.

#### G. Accuracy per Threshold Curve

To evaluate how methods behave in relation to the increase in threshold, here we plot the attack accuracy with the increase of  $th$  (Figures 6 and 7). These plots reveal an even clearer difference of behavior for the same method when attacked with either  $L_0$  or  $L_\infty$  types of attacks. It shows that the curve inclination itself is different. Therefore,  $L_0$  and  $L_\infty$  attacks scale differently.

From the figures, two classes of curves can be seen. CapsNet behave on a class of its own while the other networks behave similarly. CapsNet, which has a completely different architecture with dynamic routing, shows that a very different robustness behavior is achieved.

To assess the quality of the algorithms in relation to their curves, the Area Under the Curve (AUC) is calculated by the

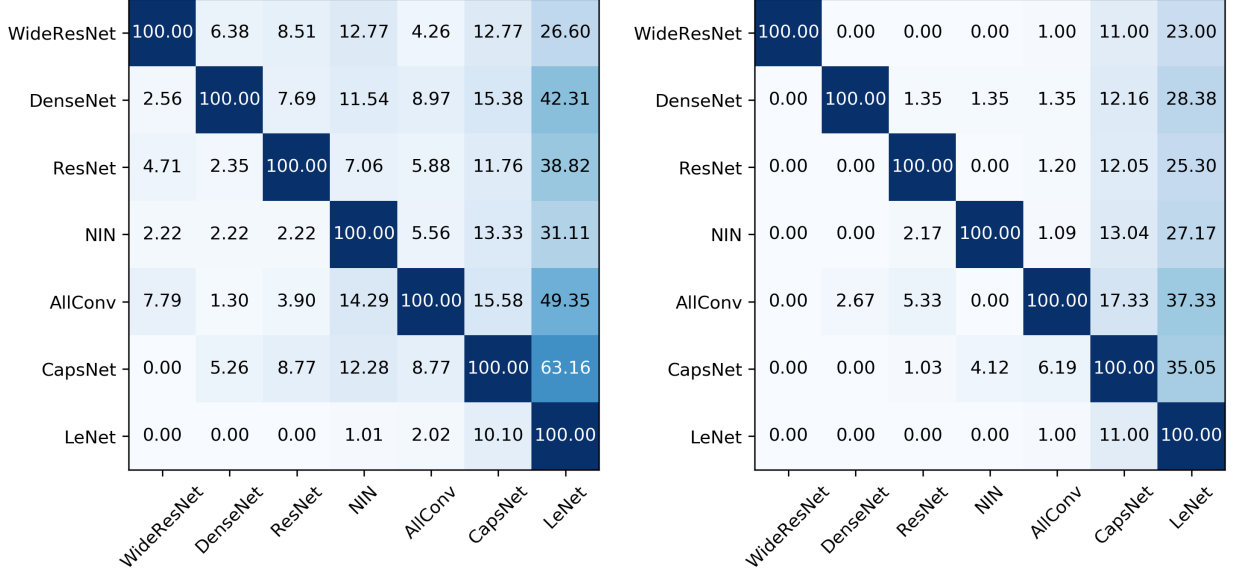


Fig. 4. Accuracy of adversarial samples when transferring from the a given source model (row) to a target model (column) for both  $L_\infty$  black-box attacks (left) and  $L_0$  black-box attacks (right). The source of the adversarial samples are on the y-axis with the target model on the x-axis. The adversarial samples were acquired from 100 original images attacked with  $th$  varying mostly from one to ten. Although most of the images were attacked with  $th < 10$ , the value of  $th$  was chosen to be the minimum value capable of attacking a given image. The maximum value of  $th$  is set to 127.

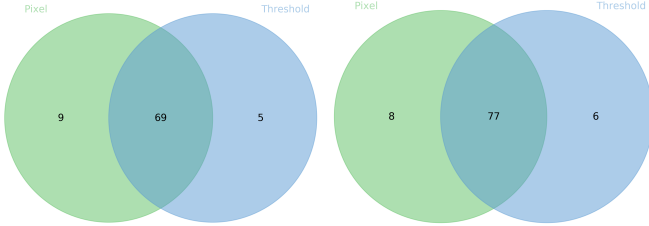


Fig. 5. Distribution of adversarial samples found on DenseNet (left) and ResNet (right) using  $th = 10$  with both few-pixel ( $L_0$ ) and threshold ( $L_\infty$ ) attacks.

trapezoidal rule defined as:

$$AUC = \Delta n_a \left( \frac{th_1}{2} + th_2 + th_3 + \dots + th_{n-1} + \frac{th_n}{2} \right),$$

where  $n_a$  is the number of images attacked and  $th_1, \dots, th_n$  are different values of  $th$  threshold for a maximum of  $n = 127$ .

Table VI shows a quantitative evaluation of Figures 6 and 7 by calculating the AUC. There is no network which is robust in both attacks. CapsNet is the most robust DNN for  $L_0$  attacks while AllConv wins while being followed closely by other DNNs for  $L_\infty$ . Although requiring a lot more resources to be drawn, the curves here result in the same conclusion achieved by Table IV. Therefore, the previous results are a good approximation of the behavior in a timely manner.

## VI. AMOUNT OF PERTURBATION

The objective of this paper is not to propose better or more effective attacking methods but rather to propose an assessment methodology and its related duality conjecture (the

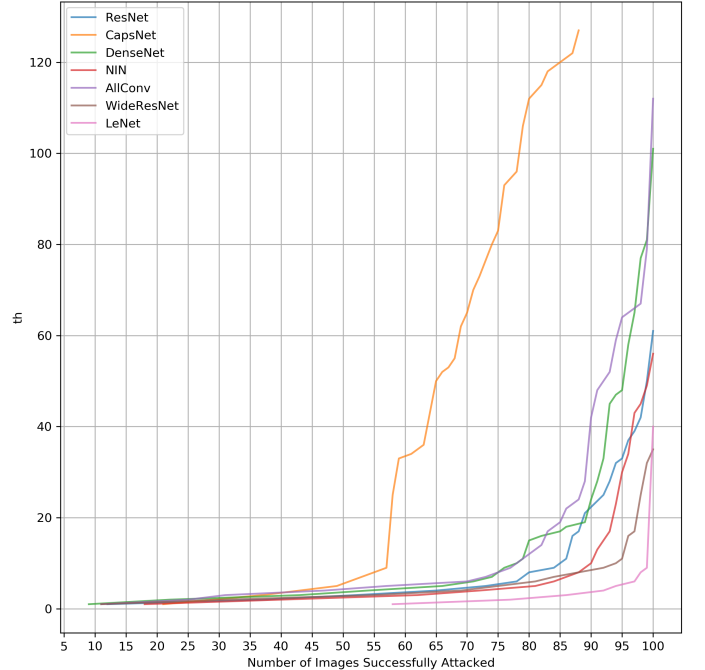


Fig. 6. Attack accuracy per  $th$  for  $L_0$  attack.

necessity of evaluating both  $L_0$  and  $L_\infty$  attacks because some methods are biased towards one of them).

Having said that, the proposed Threshold Attack in the assessment methodology is more accurate while requiring less amount of perturbation (Table VII). In fact, the proposed method needs less perturbation than the One-Pixel attack (only circa 12% of the amount of perturbation of the One-Pixel



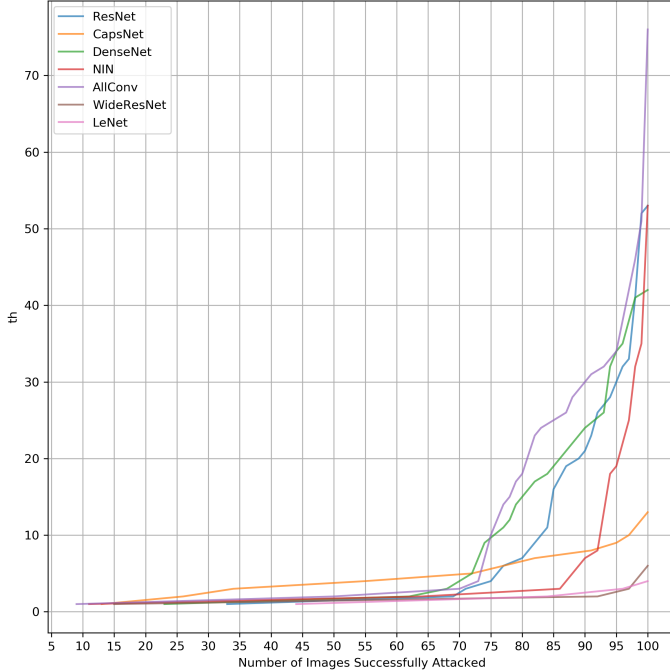


Fig. 7. Attack accuracy per  $th$  for  $L_\infty$  attack.

Model (Accuracy)	AUC for $L_0$ Attack
WideResNet (95.12)	425.0
DenseNet (94.54)	989.5
ResNet (92.67)	674.0
NIN (90.87)	528.0
AllConv (88.46)	1123.5
CapsNet (79.03)	<b>2493.0</b>
LeNet (73.57)	137.5

---

Model (Accuracy)	AUC for $L_\infty$ Attack
WideResNet (95.12)	141.5
DenseNet (94.54)	696.0
ResNet (92.67)	575.5
NIN (90.87)	364.0
AllConv (88.46)	<b>849.0</b>
CapsNet (79.03)	404.5
LeNet (73.57)	104.0

TABLE VI

AREA UNDER THE CURVE (AUC) FOR BOTH FEW-PIXEL ( $L_0$ ) AND THRESHOLD ( $L_\infty$ ) BLACK-BOX ATTACKS. LEFT COLUMN SHOW THE MODEL ATTACKED WITH THE CLASSIFICATION ACCURACY BETWEEN BRACKETS. THE ATTACK IS PERFORMED OVER 100 RANDOM SAMPLES USING THE CMA-ES OPTIMIZATION ALGORITHM.

Attack is required for  $th = 1$ ) which was already considered one of the most extreme attacks needing less perturbation to fool DNNs. This sets up an even lower threshold to the perturbation necessary to fool DNNs.

Moreover, since a  $th = 5$  is enough to achieve around 70% accuracy in many settings, this suggests that achieving 100% attack accuracy may depend more on a few samples which are harder to attack, such as samples far away from the decision boundary. Consequently, the focus on 100% attack accuracy rather than the amount of threshold, might give preference to methods which set a couple of input projections far away from others (i.e., making some input projections far away enough to

make them harder to attack), without improving the accuracy overall.

## VII. CONCLUSIONS

This work proposed a model agnostic dual quality assessment for adversarial machine learning. By analyzing the state-of-the-art models as well as arguably the current hardest defenses, it was possible to (a) show that robustness to  $L_0$  and  $L_\infty$  differ greatly and therefore duality should be taken into consideration, (b) verify that current methods and defenses in general are vulnerable even for  $L_0$  and  $L_\infty$  black-box attacks of low  $th$  and (c) validate the dual quality assessment with robustness level as a good and efficient approximation to the full accuracy per threshold curve. Moreover, we have shown that a transferability of low  $th$  adversarial samples fail to give good approximation of attacks but enables an extremely fast qualitative evaluation. Interestingly, a by-product of the evaluation here is the proposal of a novel  $L_\infty$  black-box attack based on CMA-ES. The proposed method were shown to require surprisingly less amount of perturbation, requiring only circa 12% of the amount of perturbation used by the One-Pixel Attack while achieving similar accuracy.

Thus, this paper analyze the robustness of current DNNs and defenses as well as walks towards a better evaluation of robustness by elucidating the problems as well as proposing solutions to them. Hopefully, the proposed dual quality assessment and analysis on current DNNs' robustness will aid the development of more robust DNNs and hybrids alike.

## ACKNOWLEDGMENTS

This work was supported by JST, ACT-I Grant Number JP-50166, Japan. Additionally, we would like to thank Prof. Junichi Murata for the kind support without which it would not be possible to conduct this research.

## REFERENCES

- [1] C. e. a. Szegedy, "Intriguing properties of neural networks," in *In ICLR*. Citeseer, 2014.
- [2] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 427–436.
- [3] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 86–94.
- [4] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, "Adversarial patch," *arXiv preprint arXiv:1712.09665*, 2017.
- [5] J. Su, D. V. Vargas, and S. Kouichi, "One pixel attack for fooling deep neural networks," *arXiv preprint arXiv:1710.08864*, 2017.
- [6] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," *arXiv preprint arXiv:1607.02533*, 2016.
- [7] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2016, pp. 1528–1540.
- [8] A. Athalye and I. Sutskever, "Synthesizing robust adversarial examples," in *ICML*, 2018.
- [9] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *2016 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2016, pp. 582–597.
- [10] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2017, pp. 39–57.

Attack Name	ResNet		CapsNet		AT		FS	
	Acc.	$L_2$ Score	Acc.	$L_2$ Score	Acc.	$L_2$ Score	Acc.	$L_2$ Score
Fast Gradient Method [11]	77	2586.77	94	2594.83	-	-	75	2601.25
Basic Iterative [6]	100	3476.71	100	3445.87	100	3459.23	100	3426.23
Carlini and Wagner [10]	95	2487.60	48	2750.75	39	2274.21	93	2551.54
Deep Fool [33]	43	2799.44	64	2599.08	23	2232.59	38	2646.84
Newton Fool [34]	47	2748.40	75	2525.77	57	2224.21	39	2817.67
Virtual Adversarial Methodv [35]	39	2777.07	43	2786.98	34	2738.42	10	2016.27
Few-Pixel Attack with th=1	10	271.81	14	307.75	22	264.23	17	247.74
Few-Pixel Attack with th=3	48	434.01	34	531.78	52	488.92	49	446.29
Few-Pixel Attack with th=5	72	527.88	45	660.61	66	622.65	69	551.29
Few-Pixel Attack with th=10	82	656.75	62	790.81	86	787.24	78	677.31
Threshold Attack with th=1	28	39.16	9	39.16	3	38.50	26	39.17
Threshold Attack with th=3	78	124.55	38	129.46	12	125.73	63	124.48
Threshold Attack with th=5	78	209.24	74	224.55	25	218.87	66	208.81
Threshold Attack with th=10	83	402.93	98	443.38	57	440.96	74	399.77

TABLE VII

COMPARISON OF THE  $L_0$  (FEW-PIXEL) ATTACK AND THE PROPOSED  $L_\infty$  (THRESHOLD) ATTACK USED IN THE DUAL QUALITY ASSESSMENT AND THEIR COMPARISON WITH OTHER METHODS FROM THE LITERATURE. NOTICE THAT THE RESULTS HERE WERE DRAWN BY ATTACKING A DIFFERENT SET OF SAMPLES FROM PREVIOUS ATTACKS AND THEREFORE THE ACCURACY RESULTS MAY DIFFER SLIGHTLY FROM PREVIOUS TABLES.

- [11] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.
- [12] R. Huang, B. Xu, D. Schuurmans, and C. Szepesvári, "Learning with a strong adversary," *arXiv preprint arXiv:1511.03034*, 2015.
- [13] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *ICLR*, 2018.
- [14] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," in *ICLR*, 2018.
- [15] X. Ma, B. Li, Y. Wang, S. M. Erfani, S. Wijewickrema, G. Schoenebeck, D. Song, M. E. Houle, and J. Bailey, "Characterizing adversarial subspaces using local intrinsic dimensionality," *arXiv preprint arXiv:1801.02613*, 2018.
- [16] C. Guo, M. Rana, M. Cisse, and L. van der Maaten, "Countering adversarial images using input transformations," in *ICLR*, 2018.
- [17] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in *ICML*, 2018.
- [18] J. Uesato, B. O'Donoghue, P. Kohli, and A. Oord, "Adversarial risk and the dangers of evaluating against weak attacks," in *International Conference on Machine Learning*, 2018, pp. 5032–5041.
- [19] K. Grosse, P. Manoharan, N. Papernot, M. Backes, and P. McDaniel, "On the (statistical) detection of adversarial examples," *arXiv preprint arXiv:1702.06280*, 2017.
- [20] W. Xu, D. Evans, and Y. Qi, "Feature squeezing: Detecting adversarial examples in deep neural networks," in *Network and Distributed Systems Security Symposium (NDSS)*, 2018.
- [21] N. Carlini and D. Wagner, "Adversarial examples are not easily detected: Bypassing ten detection methods," in *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*. ACM, 2017, pp. 3–14.
- [22] —, "Magnet and" efficient defenses against adversarial attacks" are not robust to adversarial examples," *arXiv preprint arXiv:1711.08478*, 2017.
- [23] D. V. Vargas and J. Su, "Understanding the one-pixel attack: Propagation maps and locality analysis," *arXiv preprint arXiv:1902.02947*, 2019.
- [24] N. Hansen, S. D. Müller, and P. Koumoutsakos, "Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (cma-es)," *Evolutionary computation*, vol. 11, no. 1, pp. 1–18, 2003.
- [25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
- [26] S. Sabour, N. Frosst, and G. E. Hinton, "Dynamic routing between capsules," in *Advances in neural information processing systems*, 2017, pp. 3856–3866.
- [27] R. Storn and K. Price, "Differential evolution—a simple and efficient heuristic for global optimization over continuous spaces," *Journal of global optimization*, vol. 11, no. 4, pp. 341–359, 1997.
- [28] S. Zagoruyko and N. Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.
- [29] F. Iandola, M. Moskewicz, S. Karayev, R. Girshick, T. Darrell, and K. Keutzer, "Densenet: Implementing efficient convnet descriptor pyramids," *arXiv preprint arXiv:1404.1869*, 2014.
- [30] M. Lin, Q. Chen, and S. Yan, "Network in network," *arXiv preprint arXiv:1312.4400*, 2013.
- [31] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," *arXiv preprint arXiv:1412.6806*, 2014.
- [32] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner *et al.*, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [33] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2574–2582.
- [34] U. Jang, X. Wu, and S. Jha, "Objective metrics and gradient descent algorithms for adversarial examples in machine learning," in *Proceedings of the 33rd Annual Computer Security Applications Conference*. ACM, 2017, pp. 262–277.
- [35] T. Miyato, S.-i. Maeda, M. Koyama, K. Nakae, and S. Ishii, "Distributional smoothing with virtual adversarial training," *arXiv preprint arXiv:1507.00677*, 2015.