**RoBIC: A BENCHMARK SUITE FOR ASSESSING CLASSIFIERS ROBUSTNESS**

*Thibault Maho, Benoît Bonnet, Teddy Furon† and Erwan Le Merrer*

Univ. Rennes, Inria, CNRS, IRISA, Rennes France

**ABSTRACT**

Many defenses have emerged with the development of adversarial attacks. Models must be objectively evaluated accordingly. This paper systematically tackles this concern by proposing a new parameter-free benchmark we coin RoBIC. RoBIC fairly evaluates the robustness of image classifiers using a new half-distortion measure. It gauges the robustness of the network against white and black box attacks, independently of its accuracy. RoBIC is faster than the other available benchmarks. We present the significant differences in the robustness of 16 recent models as assessed by RoBIC.

We make this benchmark publicly available for use and contribution at https://gitlab.inria.fr/tmaho/robustness_benchmark.

*Index Terms— Benchmark, adversarial examples, model robustness, half-distortion measure.*

1. **INTRODUCTION**

Deep learning models are vulnerable to adversarial perturbations. This is especially true in image classification in computer vision. This weakness is unfortunately undermining the development of ‘Artificial Intelligence’. In particular, adversarial attacks are a serious threat for security oriented applications. Attackers willing to bypass security countermeasures might use the deep learning models as the weakest link. A deluge of research papers now propose defenses to block such an attacker, and adaptive attacks against these defenses. This is an endless arms race, and systematic benchmarks to evaluate the state of the threat are greatly required.

It is currently extremely difficult to have a clear view on what is truly working in this domain. The cliché is that no two papers report the same statistics for the same attack against the same model over the same image set. This is mostly due to that an attack is an algorithm with many parameters. Its power is indeed highly dependent of these parameters. These values are rarely specified in research papers.

There exist benchmarks in the literature, such as ARES [1], RobustBench [2], RobustVision [3], ADBD [4]. They aim at providing a better understanding of the robustness of image classifiers. Yet, they fall short because their slowness prevents them from tackling large image dataset like ImageNet. They only operate on CIFAR-10 or MNIST. Also, they resort to attacks which are not all state-of-the-art.

This paper proposes RoBIC, to consider these concerns and develop a benchmark tool to measure the robustness of image classifiers in a modern setup.

2. **DIFFICULTIES**

This sections explains the difficulties for setting up a benchmark measuring the robustness of image classifiers.

2.1. Notation

An attack is a process forging an image $I_a = \mathcal{A}(I_o, M, \Pi)$, where $I_o$ is the original image, $M$ is the target model, and $\Pi$ is a set of attack parameters. The ground truth label of $I_o$ is denoted by $y_o$. The boolean function $1(I_a, y_o) = [M(I_a) \neq y_o]$ tells whether the attack deludes classifier $M$ in the untargeted attack scenario: the prediction $M(I_a)$ is not the ground truth. The distortion between $I_o$ and $I_a$ is denoted by $d(I_o, I_a)$.

Some statistics like the probability of success and the average distortion are extracted from the adversarial images forged from the test set. They depend on the attack $\mathcal{A}$ and its set of parameters $\Pi$. Therefore, it can not play the role of a measure of robustness of a given model. The first difficulty is to get rid off the impact of parameters $\Pi$.

2.2. The best effort mode

The parameters $\Pi$ have a huge impact on the power of an attack. For instance, some attacks like FGSM [5], I-FGSM [6], PGD [6] are distortion constrained in the sense that $\Pi$ is strongly connected to a distortion budget. If this budget is small, the probability of the success of the attack is small. If it is large, this probability is close to 1 but the distortion is too big. Hence, it is hard to find the best setting to make these attacks competitive. Our strategy, so-called ‘best effort mode’, reveals the intrinsic power of an attack by finding the best setting for any image: $I_a = \mathcal{A}(I_o, M, \Pi^*)$ with

$$
\Pi^* = \arg \min_{\Pi:1(\mathcal{A}(I_o,M,\Pi),y_o)=1} d(\mathcal{A}(I_o,M,\Pi,I_o)).
$$

The best effort mode makes the measurement of the robustness independent from an arbitrary global setting $\Pi$. Yet, it is

---

†Thanks to ANR and AID for funding Chaire IA SAIDA (ANR_20-CHIA-0011-01).
costly in terms of computations. Attacks with few parameters are preferred since the search space is smaller.

2.3. Worst case attacks
A second difficulty is to make the robustness score independent of the attack. Ideally, we would like to know the worst case attack to certify the robustness of a model. An option proposed by benchmarks RobustVision [3] and ARES [1] is to consider a set of $J = 11$ attacks as outlined in table 1. This is again costly as each image of the test set has to be attacked $J$ times. Yet, a benchmark happens to be useful if it is fast enough so to assess the robustness of many models. The best effort mode over an ensemble of attacks is out of reach. This is the reason why we need to focus on fast worst case attacks in the sense that they achieve their best effort mode within limited complexity. Section 4 focuses on these attacks.

2.4. The choice of the metric
The game between attack $A$ and model $M$ over the test set is summarized by the operating characteristic $D \rightarrow P(D)$ relating the distortion $D$ and the probability of success $P(D)$:

$$P(D) := n^{-1} \sum_{i:d(I_{a,i}, I_{o,i}) \leq D} \mathbb{1}(I_{a,i}, y_{o,i}).$$

(2)

In other words, $P(D)$ is the fraction of images that the attack succeeded to hack within a distortion budget $D$. Many benchmarks gauge the robustness by $P(D_b)$ at an arbitrary distortion $D_b$, e.g. RobustBench [2] score is $P(D = 0.5)$. This measure is pointwise and dependent on $\eta(0)$.

3. THE BENCHMARK
This section justifies the recommendations made in our benchmark and defines the measure of robustness.

Pixel domain. Our benchmark is dedicated to image classification. As a consequence, the distortion is defined on the pixel domain: An image $I$ is defined in the space $[0, 255]^n$ with $n = 3RC$ pixels for 3 color channels, $R$ rows and $L$ columns. Most papers in the field measure distortion after the transformation of the image in a tensor $x \in \mathcal{X}^n$. This is a mistake preventing a fair comparison: for most models $\mathcal{X} = [0, 1]$, but for some others $\mathcal{X} = [-1, 1]$ or $\mathcal{X} = [-3, 3]$.

We outline that an adversarial image is above all an interpretable: if $I_{a,i} = I_{o,i} + \epsilon_i$, $\forall i \in [1, n]$, then $d(I_{o,i}) = \epsilon$.

It is easily translated into a PNSR as image processing professionals do: \[ \text{PSNR} = 48.13 - 20 \log_{10}(d(I_{o,i})) \text{ dB}. \] Adversarial perturbations usually spread all over the image and have small amplitude like in invisible watermarking. This is a case where measures based on $\ell_2$ norm remain good indicators of the quality. A perceptual similarity is obviously better, but more complex and less interpretable.

Test set. The input of the model is a natural and large image. Assessing the robustness of models on specific dataset like MNIST (almost black and white), or on tiny images like CIFAR does not reflect the complexity of the problem. Our benchmark considers natural images of at least $224 \times 224$ pixels as provided in ImageNet.

Measure of robustness. Let us define the accuracy function $\eta(D) := 1 - P(D)$. The value $\eta(0)$ is the classical accuracy of the model over original images. Function $\eta(D)$ is by construction non increasing and should converge to 0 as the distortion $D$ increases. After observing many accuracy functions $\eta$ for different models and attacks, we notice that they share the same prototype:

$$\eta(D) = \eta(0) e^{-\lambda D} \quad \text{with } \lambda \in \mathbb{R}^+. \quad (3)$$

Like in nuclear physics, we define the half-distortion $D_{1/2}$ as the distortion needed to reduce to half the initial accuracy:

$$\eta(D_{1/2}) = \eta(0)/2, \quad D_{1/2} = \lambda^{-1} \log(2). \quad (4)$$

This approximation is verified experimentally with an average coefficient of determination $R^2$ of 99%. The half-distortion $D_{1/2}$ will be the keystone of the proposed metric of robustness. A model is then characterized by three separated concepts: its generalization ability $\eta(0)$ and its robustnesses $D_{1/2}$ against black-box and white-box attacks.

4. FAST ATTACKS
The recent trend in adversarial examples is to design fast attacks with state-of-the-art performances.

4.1. Fast black-box attacks
In the black-box decision based setup, the attacker can query a model and observes the predicted class. The complexity of the attack is gauged by the number of queries $K$ needed to find an adversarial image of low distortion.

There has been a huge improvement on the amount of queries recently. Brendel et al. report in the order of one million of queries for one image in one of the first decision based black-box BA [7, Fig. 6]. Then, the order of magnitude went down to tens of thousands [8, Fig. 4] [9, Fig. 5] and even some thousands in [10, Fig. 2]. Current benchmarks use others black-box attacks, which are either decision-based (Square Attack [11] in RobustBench [2] is score-based), or not state-of-the-art (like Gaussian noise in RobustVision [3], or BA [7] in ARES [1]).
SurFree [12] and RayS [4] are the only decision-based papers with less than one thousand of calls on ImageNet. Yet, RayS [4] is designed to minimize the $\ell_\infty$ distortion, whereas SurFree [12] targets $\ell_2$. Sect. 5 investigates which attack is the best candidate for a fast benchmark.

### 4.2. Fast white-box attacks


Again, the need for powerful but fast attacks is of utmost importance for a practical benchmark. A promising attack is BP [15] designed for low complexity budget. Its first stage finds an adversarial example as quickly as possible. It is nothing more than a gradient descent of the loss $L$ with acceleration. At iteration $t + 1$:

$$I_a^{(t+1)} = I_a^{(t)} - \alpha \gamma(t + 1) \eta \left( \nabla L(I_a^{(t)}) \right),$$  

(5)

where $I_a^{(0)} = I_o$, $\eta(x) = x/\|x\|_2$, and $\gamma(t)$ is a series of increasing values, hence the acceleration. Stage 1 finishes when $I_a^{(t+1)}$ becomes adversarial. Stage 2 aims at lowering the distortion while maintaining the image adversarial (see [15]).

We develop a variant to aggressively downsize the number of gradient computations. Parameter $\alpha$ is heuristically set up to 0.03 in [15]. This value is certainly too big for images close to the class boundary and too small for those further away. One costly option is the best effort mode which finds the best $\alpha$ thanks to a line search (see Sect. 2). We propose the following simple method inspired by DeepFool [13]. When applying (5) to the first order approximation of the loss:

$$L(I_o + p) \approx L(I_o) + p^\top \nabla L(I_o),$$

(6)

then $\eta \left( \nabla L(I_a^{(t)}) \right) = \eta(\nabla L(I_o))$ and BP cancels the loss for

$$\alpha = \frac{L(I_o)}{\|\nabla L(I_o)\|_2 \sum_{k=1}^\kappa \gamma(k)}$$

(7)

within $\kappa$ iterations. We fix $\kappa = \lfloor K/3 \rfloor$ where $K$ is the total iteration budget encompassing stages 1 and 2.

Sect. 5 compares these attacks to identify the worst case.

### 4.3. Quantization

The adversarial samples are quantified in the pixel domain to create images. The first option considers the quantization as a post-processing not interfering with the attack. The second option performs quantization at the end of any iteration. These options are tested on several black and white box attacks. The quantization will be a post-processing for white-box attacks as recommended in [16], whereas the second option give better results on black-box attacks.

### 5. EXPERIMENTS

All the attacks are run on 1000 ImageNet images from the ILSVRC2012’s validation set with size $n = 3 \times 224 \times 224$.

#### 5.1. Selecting the worst case attacks

**Black box attacks:** Figure 1 compares the evolution of the half-distortion (4) in function of the query amount for four decision-based black-box attacks: SurFree [12], RayS [4], GeoDA [10], and QEBA [9]. SurFree and RayS reach their best effort within 3000 queries, while QEBA and GeoDA do not since their $D_{1/2}$ still decrease after 5000 queries. Yet, SurFree obtains quantified adversarials with much lower distortion. Therefore, our benchmark only needs this attack. The number of queries is kept at 5000 to be sure to reach the optimal value of $D_{1/2}$.

**White box attacks:** Figure 2 compares three white-box attacks in the best effort mode: PGD [6], CW [14], and BP [15] with our trick (7). They all reach the same $D_{1/2}$ when given
a large complexity budget. Yet, BP converges faster than the others. Our benchmark uses this version of $BP$ to evaluate the white-box-robustness.

5.2. Comparison with other benchmarks

Table 1 lists several benchmarks. Most of them evaluate the robustness as the success-rate under a prescribed $\ell_2$ or $\ell_\infty$ distortion budget. But, these budgets are set arbitrarily or even not constant within the same benchmark for RobustML. Our half-distortion (4) is parameter-free. It returns an accurate, reliable and fair measurement of robustness.

Some benchmarks need many attacks to get a full vision of the robustness: ARES [1] and RobustVision [3] use 11 attacks. This is too time-consuming. On the contrary, ADBD [4] focuses on a single black-box attack, which is indeed outdated. RobustBench [2] condenses four attacks in one measure elegantly: for a given image, if the first simple attack does not succeed within the distortion budget, then the second more complex one is launched etc. The total runtime heavily depends on the distortion budget. Yet, black-box and white-box attacks use different mechanisms. Our benchmark reports a measurement for each separately.

5.3. Benchmarking models

Table 2 compares standard models from timm [31] and torchvision [32] libraries. Here are some intriguing results.

**Robustness in white box vs. black box.** One does not imply the other. Fig. 3 even shows a negative correlation. However, some models escape this rule. For instance, VGG16 is neither robust in black box nor in white box. EfficientNet AdvProp [28] follows the opposite trend. We believe that black-box robustness reveals the complexity of the borders between classes, and white-box robustness indicates how close natural images are from the borders. This highlights the importance of having two different measurements.

The importance of the training procedure. There is on average a factor 20 between the half-distrortions in white and black box. This factor drops to 4 and 10 for the models adversarially trained: ResNet50 [18], EfficientNet AdvProp [28].

Table 2 lists four EfficientNet models sharing the same architecture but different training procedures. Their accuracies are similar but there is up to a factor of 2 between the robustnesses. The same holds on the three variants of ResNet50. The gaps in accuracy and robustness are noticeable with standard models from timm [31] and torchvision [32]. It is even more visible with adversarial training from [18]: the gain in robustness is impressive but at the cost of a big drop in accuracy.

6. CONCLUSION

The paper introduces a rigorous benchmark based on a new and independent measurement of robustness: the half-distortion. RoBIC is faster than the other benchmarks. This allows to tackle larger images which is more realistic.

In addition to the accuracy, RoBIC gives the black box robustness, and white box robustness. We believe that the first indicates how far away the class boundaries lie from the images whereas the last reflects how curved are the boundaries. As the other benchmarks, two limitations hold: The network must be differentiable to run a white box attack, and deterministic to run a black box attack.
7. REFERENCES


