Advocating for Multiple Defense Strategies against Adversarial Examples

Alexandre Araujo[1,2], Laurent Meunier[1,3], Rafael Pinot[1,4], and Benjamin Negrevergne[1]

[1] PSL, Université Paris-Dauphine, Miles Team
[2] Wavestone
[3] Facebook AI Research
[4] CEA, Université Paris-Saclay
{firstname.lastname}@dauphine.psl.eu

Abstract. It has been empirically observed that defense mechanisms designed to protect neural networks against ℓ_∞ adversarial examples offer poor performance against ℓ_2 adversarial examples and vice versa. In this paper we conduct a geometrical analysis that validates this observation. Then, we provide a number of empirical insights to illustrate the effect of this phenomenon in practice. Then, we review some of the existing defense mechanism that attempts to defend against multiple attacks by mixing defense strategies. Thanks to our numerical experiments, we discuss the relevance of this method and state open questions for the adversarial examples community.

1 Introduction

Deep neural networks achieve state-of-the-art performances in a variety of domains such as natural language processing [19], image recognition [9] and speech recognition [10]. However, it has been shown that such neural networks are vulnerable to *adversarial examples*, *i.e.*, imperceptible variations of the natural examples, crafted to deliberately mislead the models [7, 3, 22]. Since their discovery, a variety of algorithms have been developed to generate adversarial examples (a.k.a. attacks), for example FGSM [8], PGD [15] and C&W [5], to mention the most popular ones.

Because it is difficult to characterize the space of visually imperceptible variations of a natural image, existing adversarial attacks use surrogates that can differ from one attack to another. For example, [8] use the ℓ_∞ norm to measure the distance between the original image and the adversarial image whereas [5] use the ℓ_2 norm. When the input dimension is low, the choice of the norm is of little importance because the ℓ_∞ and ℓ_2 balls overlap by a large margin, and the adversarial examples lie in the same space. An important insight in this paper is to observe that the overlap between the two balls diminishes exponentially quickly as the dimensionality of the input space increases. For typical image datasets with large dimensionality, the two balls are mostly disjoint. As a consequence, the ℓ_∞ and the ℓ_2 adversarial examples lie in different areas of the space,

and it explains why ℓ_∞ defense mechanisms perform poorly against ℓ_2 attacks and vice versa.

Building on this insight, we advocate for designing models that incorporate defense mechanisms against both ℓ_∞ and ℓ_2 attacks and review several ways of mixing existing defense mechanisms. In particular, we evaluate the performance of *Mixed Adversarial Training* (MAT) [8] which consists of augmenting training batches using *both* ℓ_∞ and ℓ_2 adversarial examples, and *Randomized Adversarial Training* (RAT) [20], a solution to benefit from the advantages of both ℓ_∞ adversarial training, and ℓ_2 randomized defense.

Outline of the paper. The rest of this paper is organized as follows. In Section 2, we recall the principle of existing attacks and defense mechanisms. In Section 3, we conduct a theoretical analysis to show why the ℓ_∞ defense mechanisms cannot be robust against ℓ_2 attacks and vice versa. We then corroborate this analysis with empirical results using real adversarial attacks and defense mechanisms. In Section 4, we discuss various strategies to mix defense mechanisms, conduct comparative experiments, and discuss the performance of each strategy.

2 Preliminaries on Adversarial Attacks and Defenses

Let us first consider a standard classification task with an input space $\mathcal{X} = [0,1]^d$ of dimension d, an output space $\mathcal{Y} = [K]$ and a data distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$. We assume the model f_θ has been trained to minimize the expectation over $\mathcal{D}$ of a loss function $\mathcal{L}$ as follows:

$$\min_\theta \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\mathcal{L}(f_\theta(x), y)\right]. \tag{1}$$

2.1 Adversarial attacks

Given an input-output pair $(x, y) \sim \mathcal{D}$, an *adversarial attack* is a procedure that produces a small perturbation $\tau \in \mathcal{X}$ such that $f_\theta(x + \tau) \neq y$. To find the best perturbation τ, existing attacks can adopt one of the two following strategies: (i) maximizing the loss $\mathcal{L}(f_\theta(x + \tau), y)$ under some constraint on $\|\tau\|_p$[5] (a.k.a. loss maximization); or (ii) minimizing $\|\tau\|_p$ under some constraint on the loss $\mathcal{L}(f_\theta(x + \tau), y)$ (a.k.a. perturbation minimization).

(i) Loss maximization. In this scenario, the procedure maximizes the loss objective function, under the constraint that the ℓ_p norm of the perturbation remains bounded by some value ϵ, as follows:

$$\underset{\|\tau\|_p \leq \epsilon}{\operatorname{argmax}} \mathcal{L}(f_\theta(x + \tau), y). \tag{2}$$

The typical value of ϵ depends on the norm $\|\cdot\|_p$ considered in the problem setting. In order to compare ℓ_∞ and ℓ_2 attacks of similar strength, we choose

[5] with $p \in \{0, \cdots, \infty\}$.

values of ϵ_∞ and ϵ_2 (for ℓ_∞ and ℓ_2 norms respectively) which result in ℓ_∞ and ℓ_2 balls of equivalent volumes. For the particular case of CIFAR-10, this would lead us to choose $\epsilon_\infty = 0.03$ and $\epsilon_2 = 0.8$ which correspond to the maximum values chosen empirically to avoid the generation of visually detectable perturbations. The current state-of-the-art method to solve Problem (2) is based on a projected gradient descent (PGD) [15] of radius ϵ. Given a budget ϵ, it recursively computes

$$x^{t+1} = \prod_{B_p(x,\epsilon)} \left(x^t + \alpha \underset{\delta \text{ s.t. } ||\delta||_p \leq 1}{\operatorname{argmax}} \left(\Delta^t | \delta \right) \right) \tag{3}$$

where $B_p(x, \epsilon) = \{x + \tau \text{ s.t. } \|\tau\|_p \leq \epsilon\}$, $\Delta^t = \nabla_x \mathcal{L}\left(f_\theta\left(x^t\right), y\right)$, α is a gradient step size, and $\prod_S$ is the projection operator on S. Both PGD attacks with $p = 2$, and $p = \infty$ are currently used in the literature as state-of-the-art attacks for the loss maximization problem.

(ii) Perturbation minimization. This type of procedure search for the perturbation that has the minimal ℓ_p norm, under the constraint that $\mathcal{L}(f_\theta(x + \tau), y)$ is bigger than a given bound c:

$$\underset{\mathcal{L}(f_\theta(x+\tau),y) \geq c}{\operatorname{argmin}} \|\tau\|_p. \tag{4}$$

The value of c is typically chosen depending on the loss function $\mathcal{L}$[6]. Problem (4) has been tackled in [5], leading to the following method, denoted C&W attack in the rest of the paper. It aims at solving the following Lagrangian relaxation of Problem (4):

$$\underset{\tau}{\operatorname{argmin}} \|\tau\|_p + \lambda \times g(x + \tau) \tag{5}$$

where $g(x + \tau) < 0$ if and only if $\mathcal{L}(f_\theta(x + \tau), y) \geq c$. The authors use a change of variable $\tau = \tanh(w) - x$ to ensure that $-1 \leq x + \tau \leq 1$, a binary search to optimize the constant c, and Adam or SGD to compute an approximated solution. The C&W attack is well defined both for $p = 2$, and $p = \infty$, but there is a clear empirical gap of efficiency in favor of the ℓ_2 attack.

In this paper, we focus on the *Loss Maximization* setting using the PGD attack. However we conduct some of our experiments using *Perturbation Minimization* algorithms such as C&W to capture more detailed information about the location of adversarial examples in the vector space[7].

2.2 Defense mechanisms

Adversarial Training (AT). Adversarial Training was introduced in [8] and later improved in [15] as a first defense mechanism to train robust neural networks. It consists in augmenting training batches with adversarial examples generated during the training procedure. The standard training procedure from Equation (1)

[6] For example, if $\mathcal{L}$ is the 0/1 loss, any $c > 0$ is acceptable.

[7] As it has a more flexible geometry than the *Loss Maximization* attacks.

is thus replaced by the following min max problem, where the classifier tries to minimize the expected loss under maximum perturbation of its input:

$$\min_{\theta} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[\max_{\|\tau\|_p \leq \epsilon} \mathcal{L}\left(f_\theta(x+\tau), y\right) \right]. \tag{6}$$

In the case where $p = \infty$, this technique offers good robustness against ℓ_∞ attacks [1]. AT can also be used with ℓ_2 attacks but as we will discuss in Section 3, AT with one norm offers poor protection against the other. The main weakness of Adversarial Training is its lack of formal guarantees. Despite some recent work providing great insights [21, 25], there is no worst case lower bound yet on the accuracy under attack of this method.

Noise injection mechanisms (NI). Another important technique to defend against adversarial examples is to use Noise Injection. In contrast with Adversarial Training, Noise Injection mechanisms are usually deployed after training. In a nutshell, it works as follows. At inference time, given a unlabeled sample x, the network outputs

$$\tilde{f}_\theta(x) := f_\theta(x+\eta) \quad \text{(instead of } f_\theta(x)\text{)} \tag{7}$$

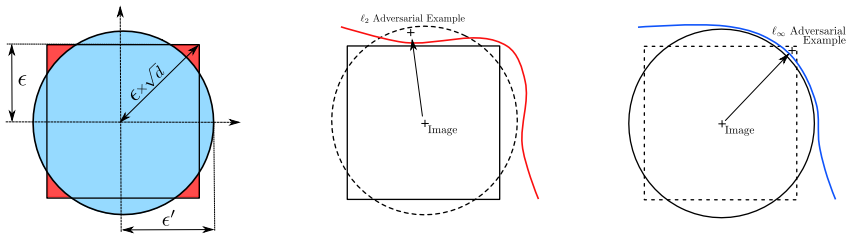
where η is a random variable on $\mathbb{R}^d$. Even though, Noise Injection is often less efficient than Adversarial Training in practice (see *e.g.*, Table 3), it benefits from strong theoretical background. In particular, recent works [13, 14], followed by [6, 18] demonstrated that noise injection from a Gaussian distribution can give provable defense against ℓ_2 adversarial attacks. In this work, besides the classical Gaussian noises already investigated in previous works, we evaluate the efficiency of Uniform distributions to defend against ℓ_2 adversarial examples.

3 No Free Lunch for Adversarial Defenses

In this Section, we show both theoretically and empirically that defenses mechanisms intending to defend against ℓ_∞ attacks cannot provide suitable defense against ℓ_2 attacks. Our reasoning is perfectly general; hence we can similarly demonstrate the reciprocal statement, but we focus on this side for simplicity.

3.1 Theoretical analysis

Let us consider a classifier f_∞ that is provably robust against adversarial examples with maximum ℓ_∞ norm of value ϵ_∞. It guarantees that for any input-output pair $(x, y) \sim \mathcal{D}$ and for any perturbation τ such that $\|\tau\|_\infty \leq \epsilon_\infty$, f_∞ is not misled by the perturbation, *i.e.*, $f_\infty(x+\tau) = f_\infty(x)$. We now focus our study on the performance of this classifier against adversarial examples bounded with a ℓ_2 norm of value ϵ_2. Using Figure 1(a), we observe that any ℓ_2 adversarial example that is also in the ℓ_∞ ball, will not fool f_∞. Conversely, if it is outside the ball, we have no guarantee.



(a) (b) (c)

Fig. 1. Left: 2D representation of the ℓ_∞ and ℓ_2 balls of respective radius ϵ and ϵ'. Middle: a classifier trained with ℓ_∞ adversarial perturbations (materialized by the red line) remains vulnerable to ℓ_2 attacks. Right: a classifier trained with ℓ_2 adversarial perturbations (materialized by the blue line) remains vulnerable to ℓ_∞ attacks.

To characterize the probability that such an ℓ_2 perturbation fools an ℓ_∞ defense mechanism in the general case (*i.e.*, any dimension d), we measure the ratio between the volume of the intersection of the ℓ_∞ ball of radius ϵ_∞ and the ℓ_2 ball of radius ϵ_2. As Theorem 1 shows, this ratio depends on the dimensionality d of the input vector x, and rapidly converges to zero when d increases. Therefore a defense mechanism that protects against all ℓ_∞ bounded adversarial examples is unlikely to be efficient against ℓ_2 attacks.

Theorem 1 (Probability of the intersection goes to 0).
Let $B_{2,d}(\epsilon) := \left\{\tau \in \mathbb{R}^d \ s.t \ \|\tau\|_2 \leq \epsilon\right\}$ *and* $B_{\infty,d}(\epsilon') := \left\{\tau \in \mathbb{R}^d \ s.t \ \|\tau\|_\infty \leq \epsilon'\right\}$. *If for all* d, *we select* ϵ *and* ϵ*' such that* $\mathrm{Vol}\left(B_{2,d}(\epsilon)\right) = \mathrm{Vol}\left(B_{\infty,d}(\epsilon')\right)$, *then*

$$\frac{\mathrm{Vol}\left(B_{2,d}(\epsilon) \bigcap B_{\infty,d}(\epsilon')\right)}{\mathrm{Vol}\left(B_{\infty,d}(\epsilon')\right)} \to 0 \ \textit{when} \ d \to \infty.$$

Proof. Without loss of generality, let us fix $\epsilon = 1$. One can show that for all d,

$$\mathrm{Vol}\left(B_{2,d}\left(\frac{2}{\sqrt{\pi}}\Gamma\left(\frac{d}{2}+1\right)^{1/d}\right)\right) = \mathrm{Vol}\left(B_{\infty,d}\left(1\right)\right) \tag{8}$$

where Γ is the gamma function. Let us denote

$$r_2(d) = \frac{2}{\sqrt{\pi}}\Gamma\left(\frac{d}{2}+1\right)^{1/d}. \tag{9}$$

Then, thanks to Stirling's formula

$$r_2(d) \sim \sqrt{\frac{2}{\pi e}} d^{1/2}. \tag{10}$$

Finally, if we denote $\mathcal{U}_S$, the uniform distribution on set S, by using Hoeffding inequality between Equation 14 and 15, we get:

$$\frac{\text{Vol}(B_{2,d}(r_2(d)) \bigcap B_{\infty,d}(1))}{\text{Vol}(B_{\infty,d}(1))} \tag{11}$$

$$= \mathbb{P}_{x \sim \mathcal{U}_{B_{\infty,d}(1)}} \left[x \in B_{2,d}(r_2(d)) \right] \tag{12}$$

$$= \mathbb{P}_{x \sim \mathcal{U}_{B_{\infty,d}(1)}} \left[\sum_{i=1}^{d} |x_i|^2 \leq r_2^2(d) \right] \tag{13}$$

$$\leq \exp\left\{ -d^{-1} \left(r_2^2(d) - d\mathbb{E}|x_1|^2 \right)^2 \right\} \tag{14}$$

$$\leq \exp\left\{ -\left(\frac{2}{\pi e} - \frac{1}{3} \right)^2 d + o(d) \right\}. \tag{15}$$

Then the ratio between the volume of the intersection of the ball and the volume of the ball converges towards 0 when d goes to ∞. □

Theorem 1 states that, when d is large enough, ℓ_2 bounded perturbations have a null probability of being also in the ℓ_∞ ball of the same volume. As a consequence, for any value of d that is large enough, a defense mechanism that offers full protection against ℓ_∞ adversarial examples is not guaranteed to offer any protection against ℓ_2 attacks[8].

Table 1. Bounds of Theorem 1 on the volume of the intersection of ℓ_2 and ℓ_∞ balls at equal volume for typical image classification datasets. When $d = 2$, the bound is $10^{-0.009} \approx 0.98$.

Dataset	**Dim. (d)**	**Vol. of the intersection**
–	2	$10^{-0.009}$ (≈ 0.98)
MNIST	784	10^{-144}
CIFAR	3072	10^{-578}
ImageNet	150528	10^{-28946}

Note that this result defeats the 2-dimensional intuition: if we consider a 2 dimensional problem setting, the ℓ_∞ and the ℓ_2 balls have an important overlap (as illustrated in Figure 1(a)) and the probability of sampling at the intersection of the two balls is bounded by approximately 98%. However, as we increase the dimensionality d, this probability quickly becomes negligible, even for very simple image datasets such as MNIST. An instantiation of the bound for classical image datasets is presented in Table 1. The probability of sampling at the intersection of the ℓ_∞ and ℓ_2 balls is close to zero for any realistic image setting. In large dimensions, the volume of the corner of the ℓ_∞ ball is much bigger than it appears in Figure 1(a).

[8] Th. 1 can easily be extended to any two balls with different norms. For clarity, we restrict to the case of ℓ_∞ and ℓ_2 norms.

3.2 No Free Lunch in Practice

Our theoretical analysis shows that if adversarial examples were uniformly distributed in a high-dimensional space, then any mechanism that perfectly defends against ℓ_∞ adversarial examples has a null probability of protecting against ℓ_2-bounded adversarial attacks. Although existing defense mechanisms do not necessarily assume such a distribution of adversarial examples, we demonstrate that whatever distribution they use, it offers no favorable bias with respect to the result of Theorem 1. As we discussed in Section 2, there are two distinct attack settings: loss maximization (PGD) and perturbation minimization (C&W). Our analysis is mainly focusing on loss maximization attacks. However, these attacks have a very strict geometry[9]. This is why, to present a deeper analysis of the behavior of adversarial attacks and defenses, we also present a set of experiments that use perturbation minimization attacks.

Table 2. Average norms of PGD-ℓ_2 and PGD-ℓ_∞ adversarial examples with and without ℓ_∞ adversarial training on CIFAR-10 ($d = 3072$).

	Attack PGD-ℓ_2		Attack PGD-ℓ_∞	
	Unprotected	AT-ℓ_∞	Unprotected	AT-ℓ_2
Average ℓ_2 norm	0.830	0.830	1.400	1.640
Average ℓ_∞ norm	0.075	0.200	0.031	0.031

Adversarial training vs. loss maximization attacks To demonstrate that ℓ_∞ adversarial training is not robust against PGD-ℓ_2 attacks we measure the evolution of ℓ_2 norm of adversarial examples generated with PGD-ℓ_∞ between an unprotected model and a model trained with AT-ℓ_∞, *i.e.*, AT where adversarial examples are generated with PGD-ℓ_∞[10]. Results are presented in Table 2. [11]

The analysis is unambiguous: the average ℓ_∞ norm of a bounded ℓ_2 perturbation more than double between an unprotected model and a model trained with AT PGD-ℓ_∞. This phenomenon perfectly reflects the illustration of Figure 1 (c). The attack will generate an adversarial example on the corner of the ℓ_∞ ball thus increasing the ℓ_∞ norm while maintaining the same ℓ_2 norm. We can observe the same phenomenon with AT-ℓ_2 against PGD-ℓ_∞ attack (see Figure 1 (b) and Table 2). PGD-ℓ_∞ attack increases the ℓ_2 norm while maintaining the same ℓ_∞ perturbation thus generating the perturbation in the upper area.

[9] Due to the projection operator, all PGD attacks saturate the constraint, which makes them all lies in a very small part of the ball.

[10] To do so, we use the same experimental setting as in Section 4 with ϵ_∞ and ϵ_2 such that the volumes of the two balls are equal.

[11] All experiments in this section are conducted on CIFAR-10, and the experimental setting is fully detailed in Section 4.1.

As a consequence, we cannot expect adversarial training ℓ_∞ to offer any guaranteed protection against ℓ_2 adversarial examples .

Adversarial training vs. perturbation minimization attacks. To better capture the behavior of ℓ_2 adversarial examples, we now study the performances of an ℓ_2 perturbation minimization attack (C&W) with and without AT-ℓ_∞. It allows us to understand in which area C&W discovers adversarial examples and the impact of AT-ℓ_∞. In high dimensions, the red corners (see Figure 1 (a)) are very far away from the ℓ_2 ball. Therefore, we hypothesize that a large proportion of the ℓ_2 adversarial examples will remain unprotected. To validate this assumption, we measure the proportion of adversarial examples inside of the ℓ_2 ball before and after ℓ_∞ adversarial training. The results are presented in Figure 2 (left: without adversarial training, right: with adversarial training).

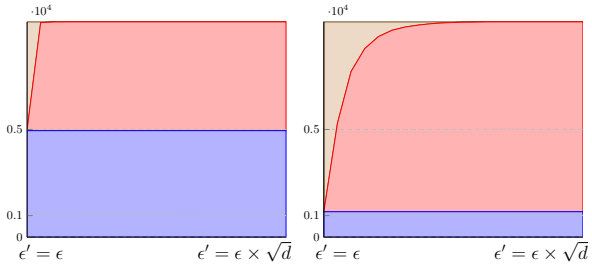


Fig. 2. Comparison of the number of adversarial examples found by C&W, inside the ℓ_∞ ball (lower, blue area), outside the ℓ_∞ ball but inside the ℓ_2 ball (middle, red area) and outside the ℓ_2 ball (upper gray area). ϵ is set to 0.3 and ϵ' varies along the x-axis. Left: without adversarial training, right: with adversarial training. Most adversarial examples have shifted from the ℓ_∞ ball to the cap of the ℓ_2 ball, but remain at the same ℓ_2 distance from the original example.

On both charts, the blue area represents the proportion of adversarial examples that are inside the ℓ_∞ ball. The red area represents the adversarial examples that are outside the ℓ_∞ ball but still inside the ℓ_2 ball (valid ℓ_2 adversarial examples). Finally, the brown-beige area represents the adversarial examples that are beyond the ℓ_2 bound. The radius ϵ' of the ℓ_2 ball varies along the x-axis from ϵ' to $\epsilon'\sqrt{d}$. On the left chart (without adversarial training) most ℓ_2 adversarial examples generated by C&W are inside both balls. On the right chart most of the adversarial examples have been shifted out the ℓ_∞ ball. This is the expected consequence of ℓ_∞ adversarial training. However, these adversarial examples remain in the ℓ_2 ball, *i.e.*, they are in the cap of the ℓ_2 ball. These examples are equally good from the ℓ_2 perspective. This means that even after adversarial training, it is still easy to find good ℓ_2 adversarial examples, making the ℓ_2 robustness of AT-ℓ_∞ almost null.

4 Reviewing Defenses Against Multiple Attacks

Table 3. This table shows a comprehensive list of results consisting of the accuracy of several defense mechanisms against ℓ_2 and ℓ_∞ attacks. This table main objective is to compare the overall performance of 'single' norm defense mechanisms (AT and NI presented in the Section 2.2) against mixed norms defense mechanisms (MAT & RAT mixed defenses presented in Section 4).

	Baseline	**AT**		**MAT**		**NI**		**RAT-ℓ_∞**		**RAT-ℓ_2**	
	–	ℓ_∞	ℓ_2	Max	Rand	$\mathcal{N}$	$\mathcal{U}$	$\mathcal{N}$	$\mathcal{U}$	$\mathcal{N}$	$\mathcal{U}$
Natural	0.94	0.85	0.85	0.80	0.80	0.79	0.87	0.74	0.80	0.79	0.87
PGD-ℓ_∞	0.00	0.43	0.37	0.37	0.40	0.23	0.22	0.35	0.40	0.23	0.22
PGD-ℓ_2	0.00	0.37	0.52	0.50	0.55	0.34	0.36	0.43	0.39	0.34	0.37

Adversarial attacks have been an active topic in the machine learning community since their discovery [7, 3, 22]. Many attacks have been developed. Most of them solve a loss maximization problem with either ℓ_∞ [8, 12, 15], ℓ_2 [5, 12, 15], ℓ_1 [23] or ℓ_0 [16] surrogate norms. As we showed, these norms are really different in high dimension. Hence, defending against one norm-based attack is not sufficient to protect against another one. In order to solve this problem, we review several strategies to build defenses against multiple adversarial attacks. These strategies are based on the idea that both types of defense must be used simultaneously in order for the classifier to be protected against multiple attacks. The detailed description of the experimental setting is described in Section 4.1.

4.1 Experimental Setting

To compare the robustness provided by the different defense mechanisms, we use strong adversarial attacks and a conservative setting: the attacker has a total knowledge of the parameters of the model (white-box setting) and we only consider untargeted attacks (a misclassification from one target to any other will be considered as adversarial). To evaluate defenses based on Noise Injection, we use *Expectation Over Transformation* (EOT), the rigorous experimental protocol proposed by [2] and later used by [1, 4] to identify flawed defense mechanisms.

To attack the models, we use state-of-the-art algorithms PGD. We run PGD with 20 iterations to generate adversarial examples and with 10 iterations when it is used for adversarial training. The maximum ℓ_∞ bound is fixed to 0.031 and the maximum ℓ_2 bound is fixed to 0.83. As discussed in Section 2, we chose these values so that the ℓ_∞ and the ℓ_2 balls have similar volumes. Note that 0.83 is slightly above the values typically used in previous publications in the area, meaning the attacks are stronger, and thus more difficult to defend against.

All experiments are conducted on CIFAR-10 with the Wide-Resnet 28-10 architecture. We use the training procedure and the hyper-parameters described

in the original paper by [24]. Training time varies from 1 day (AT) to 2 days (MAT) on 4 GPUs-V100 servers.

4.2 MAT – Mixed Adversarial Training

Earlier results have shown that AT-ℓ_p improves the robustness against corresponding ℓ_p-bounded adversarial examples, and the experiments we present in this section corroborate this observation (See Table 3, column: AT). Building on this, it is natural to examine the efficiency of *Mixed Adversarial Training* (MAT) against mixed ℓ_∞ and ℓ_2 attacks. MAT is a variation of AT that uses both ℓ_∞-bounded adversarial examples and ℓ_2-bounded adversarial examples as training examples. As discussed in [23], there are several possible strategies to mix the adversarial training examples. The first strategy (MAT-Rand) consists in randomly selecting one adversarial example among the two most damaging ℓ_∞ and ℓ_2, and to use it as a training example, as described in Equation (16):

MAT-Rand :

$$\min_{\theta} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[\mathop{\mathbb{E}}_{p\sim\mathcal{U}(\{2,\infty\})} \max_{\|\tau\|_p \leq \epsilon} \mathcal{L}\left(f_\theta(x+\tau), y\right) \right]. \tag{16}$$

An alternative strategy is to systematically train the model with the most damaging adversarial example (ℓ_∞ or ℓ_2). As described in Equation (17):

MAT-Max :

$$\min_{\theta} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[\max_{p\in\{2,\infty\}} \max_{\|\tau\|_p \leq \epsilon} \mathcal{L}\left(f_\theta(x+\tau), y\right) \right]. \tag{17}$$

The accuracy of MAT-Rand and MAT-Max are reported in Table 3 (Column: MAT). As expected, we observe that MAT-Rand and MAT-Max offer better robustness both against PGD-ℓ_2 and PGD-ℓ_∞ adversarial examples than the original AT does. More generally, we can see that AT is a good strategy against loss maximization attacks, and thus it is not surprising that MAT is a good strategy against mixed loss maximization attacks. However efficient in practice, MAT (for the same reasons as AT) lacks theoretical arguments. In order to get the best of both worlds, [20] proposed to mix adversarial training with randomization.

4.3 RAT – Randomized Adversarial Training

We now examine the performance of Randomized Adversarial Training (RAT) first introduced in [20]. This technique mixes Adversarial Training with Noise Injection. The corresponding loss function is defined as follows:

$$\min_{\theta} \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[\max_{\|\tau\|_p \leq \epsilon} \mathcal{L}\left(\tilde{f}_\theta(x+\tau), y)\right) \right]. \tag{18}$$

where $\tilde{f}_\theta$ is a randomized neural network with noise injection as described in Section 2.2, and $\|\cdot\|_p$ define which kind of AT is used. For each setting, we consider two noise distributions, Gaussian and Uniform as we did with NI. We also consider two different Adversarial training AT-ℓ_∞ as well as AT-ℓ_2.

The results of RAT are reported in Table 3 (Columns: RAT-ℓ_∞ and RAT-ℓ_2). We can observe that RAT-ℓ_∞ offers the best extra robustness with both noises, which is consistent with previous experiments, since AT is generally more effective against ℓ_∞ attacks whereas NI is more effective against ℓ_2-attacks. Overall, RAT-ℓ_∞ and a noise from uniform distribution offers the best performances but is still weaker than MAT-Rand. These results are also consistent with the literature, since adversarial training (and its variants) is the best defense against adversarial examples so far.

5 Conclusion & Perspective

In this paper, we tackled the problem of protecting neural networks against multiple attacks crafted from different norms. We demonstrated and gave a geometrical interpretation to explain why most defense mechanisms can only protect against one type of attack. Then we reviewed existing strategies that mix defense mechanisms in order to build models that are robust against multiple adversarial attacks. We conduct a rigorous and full comparison of *Randomized Adversarial Training* and *Mixed Adversarial Training* as defenses against multiple attacks.

We could argue that both techniques offer benefits and limitations. We have observed that MAT offers the best empirical robustness against multiples adversarial attacks but this technique is computationally expensive which hinders its use in large-scale applications. Randomized techniques have the important advantage of providing theoretical guarantees of robustness and being computationally cheaper. However, the certificate provided by such defenses is still too small for strong attacks. Furthermore, certain Randomized defenses also suffer from the curse of dimensionality as recently shown by [11].

Although, randomized defenses based on noise injection seem limited in terms of accuracy under attack and scalability, they could be improved either by Learning the best distribution to use or by leveraging different types of randomization such as discrete randomization first proposed in [17]. We believe that these certified defenses are the best solution to ensure the robustness of classifiers deployed into real-world applications.

6 Acknowledgement

This work was granted access to the HPC resources of IDRIS under the allocation 2020-101141 made by GENCI. We would like to thank Jamal Atif, Florian Yger and Yann Chevaleyre for their valuable insights.

Bibliography

[1] A. Athalye, N. Carlini, and D. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 274–283, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

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

[3] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 387–402. Springer, 2013.

[4] N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber, D. Tsipras, I. Goodfellow, and A. Madry. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019.

[5] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57. IEEE, 2017.

[6] J. M. Cohen, E. Rosenfeld, and J. Z. Kolter. Certified adversarial robustness via randomized smoothing. *CoRR*, abs/1902.02918, 2019.

[7] A. Globerson and S. Roweis. Nightmare at test time: robust learning by feature deletion. In *Proceedings of the 23rd international conference on Machine learning*, pages 353–360, 2006.

[8] I. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[10] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, B. Kingsbury, et al. Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal processing magazine*, 29, 2012.

[11] A. Kumar, A. Levine, T. Goldstein, and S. Feizi. Curse of dimensionality on randomized smoothing for certifiable robustness. *arXiv preprint arXiv:2002.03239*, 2020.

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

[13] M. Lecuyer, V. Atlidakais, R. Geambasu, D. Hsu, and S. Jana. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 727–743, 2018.

[14] B. Li, C. Chen, W. Wang, and L. Carin. Certified adversarial robustness with additive noise. In *Advances in Neural Information Processing Systems 32*, pages 9459–9469. Curran Associates, Inc., 2019.

[15] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[16] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami. The limitations of deep learning in adversarial settings. In *Security and Privacy (EuroS&P), 2016 IEEE European Symposium on*, pages 372–387. IEEE, 2016.

[17] R. Pinot, R. Ettedgui, G. Rizk, Y. Chevaleyre, and J. Atif. Randomization matters how to defend against strong adversarial attacks. In H. D. III and A. Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7717–7727. PMLR, 13–18 Jul 2020.

[18] R. Pinot, L. Meunier, A. Araujo, H. Kashima, F. Yger, C. Gouy-Pailler, and J. Atif. Theoretical evidence for adversarial robustness through randomization. In *Advances in Neural Information Processing Systems 32*, pages 11838–11848, 2019.

[19] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAi, 2018.

[20] H. Salman, J. Li, I. Razenshteyn, P. Zhang, H. Zhang, S. Bubeck, and G. Yang. Provably robust deep learning via adversarially trained smoothed classifiers. In *Advances in Neural Information Processing Systems*, pages 11289–11300, 2019.

[21] A. Sinha, H. Namkoong, R. Volpi, and J. Duchi. Certifying some distributional robustness with principled adversarial training, 2017.

[22] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

[23] F. Tramèr and D. Boneh. Adversarial training and robustness for multiple perturbations. *arXiv preprint arXiv:1904.13000*, 2019.

[24] S. Zagoruyko and N. Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[25] H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. E. Ghaoui, and M. I. Jordan. Theoretically principled trade-off between robustness and accuracy. *arXiv preprint arXiv:1901.08573*, 2019.