Robust Neural Networks using Randomized Adversarial Training

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

This paper tackles the problem of defending a neural network against adversarial attacks crafted with different norms (in particular \(\ell_\infty\) and \(\ell_2\) bounded adversarial examples). It has been observed that defense mechanisms designed to protect against one type of attacks often offer poor performance against the other. We show that \(\ell_\infty\) defense mechanisms cannot offer good protection against \(\ell_2\) attacks and vice-versa, and we provide both theoretical and empirical insights on this phenomenon. Then, we discuss various ways of combining existing defense mechanisms in order to train neural networks robust against both types of attacks. Our experiments show that these new defense mechanisms offer better protection when attacked with both norms.

1 Introduction

Deep neural networks achieve state of the art performances in a variety of domains such as natural language processing (Radford et al., 2018), image recognition (He et al., 2016) and speech recognition (Hinton et al., 2012). However, it has been shown that such neural networks are vulnerable to adversarial examples, i.e. imperceptible variations of natural examples, crafted to deliberately mislead the models (Globerson and Roweis, 2006; Biggio et al., 2013; Szegedy et al., 2014). Since their discovery, a variety of algorithms have been developed to generate adversarial examples (a.k.a. attacks), for example FGSM (Goodfellow et al., 2015), PGD (Madry et al., 2018) and C&W (Carlini and Wagner, 2017), 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, Goodfellow et al. (2015) use the \(\ell_\infty\) norm to measure the distance between the original image and the adversarial image whereas Carlini and Wagner (2017) use the \(\ell_2\) norm. When the input dimension is low, the choice of the norm is of little importance because the \(\ell_\infty\) and \(\ell_2\) 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 increases. For typical image datasets with large dimensionality, the two balls are mostly disjoint. As a consequence, the \(\ell_\infty\)-bounded and the \(\ell_2\)-bounded adversarial examples lie in different area of the space, and it explains why \(\ell_\infty\) defense mechanisms perform poorly against \(\ell_2\) attacks and vice-versa.

We show that this insight is crucial to design defense mechanisms that are robust against both types of attacks, and we advocate for the design of models that incorporate defense mechanisms against both \(\ell_\infty\) and \(\ell_2\) attacks. Then we evaluate strategies (existing and new ones) to mix up existing defense mechanisms. In particular, we evaluate the following strategies:

(a) Mixed Adversarial Training (MAT), a training procedure inspired by Adversarial Training (Goodfellow et al., 2015). It is based on augmenting training batches using both \(\ell_\infty\) and \(\ell_2\) adversarial examples. This method defends well against both norms for PGD attacks, but fails against C&W attacks.

(b) Mixed noise injection (MNI), a technique that consists in noise injection at test time (Cohen et al., 2019; Pinot et al., 2019). We evaluate different noises and their mixture. This method defends better against C&W attacks, but does not obtain good results against PGD attacks for \(\ell_\infty\) norm.

(c) Randomized Adversarial Training (RAT), a solution to benefit from the advantages of both \(\ell_\infty\) adversarial training, and \(\ell_2\) randomized defense. As we will show, RAT offers the best trade-off between defending against PGD and C&W attacks.

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 \(\ell_\infty\) defense mechanisms cannot be robust against \(\ell_2\) attacks and vice-versa. We then corroborate this analysis with empirical results using real adversarial attacks and defense mechanisms.
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 Defense Mechanisms

Let us first consider a standard classification task with an input space $X = [0, 1]^d$ of dimension $d$, an output space $Y = [K]$ and a data distribution $D$ over $X \times Y$. We assume the model $f_0$ has been trained to minimize a loss function $L$ as follows:

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

(1)

In this paper, we consider $N$-layers neural network models, therefore the model is a composition of $N$ non-linear parametric functions $\phi_{\theta_i}$, i.e., $f_\theta = \phi_{\theta_N} \circ \ldots \circ \phi_{\theta_1}$.

2.1 Adversarial attacks

Given an input-output pair $(x,y) \sim D$, an adversarial attack is a procedure that produces a small perturbation $\tau \in X$ such that $f_\theta(x + \tau) \neq y$. To discover the damaging perturbation $\tau_x \in X$ which results in values of $\epsilon$ and $\ell_p$ norms of the perturbation remains bounded by some value $\epsilon$, one can perform the following two strategies: (i) maximizing the loss $L(f_\theta(x + \tau), y)$ under some constraint on $\|\tau\|_p$, with $\tau \in \{0, \ldots, \epsilon\}$ (a.k.a. loss maximization); or (ii) minimizing $\|\tau\|_p$ under some constraint on the loss $L(f_\theta(x + \tau), y)$ (a.k.a. perturbation minimization).

(i) Loss maximization. In this scenario, the procedure maximizes the loss objective function, under the constraint that the $\epsilon$-minimal $\ell_p$ norm of the perturbation remains bounded by some value $\epsilon$, as follows:

$$\max_{\|\tau\|_p \leq \epsilon} L(f_\theta(x + \tau), y).$$

(2)

The typical value of $\epsilon$ depends on the value $p$ of the norm $\|\cdot\|_p$ considered in the problem setting. In order to compare $\epsilon_\infty$ and $\epsilon_2$ attacks of similar strength, we choose values of $\epsilon_\infty$ and $\epsilon_2$ for $\ell_\infty$ and $\ell_2$ norms respectively) which result in $\ell_\infty$ and $\ell_2$ 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) (Madry et al., 2018) of radius $\epsilon$. Given a budget $\epsilon$, it recursively computes

$$x^{t+1} = \prod_{B_p(x,\epsilon)} \left( x^t + \alpha \arg\max_{\delta \text{ s.t. } \|\delta\|_p \leq 1} \left( \Delta^t|\delta| \right) \right),$$

(3)

where $B_p(x,\epsilon) = \{x + \tau \text{ s.t. } \|\tau\|_p \leq \epsilon\}$, $\Delta^t = \nabla_x f_\theta(x^t), y$, $\alpha$ is a gradient step size, and $\prod_{B_p}$ 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 procedures search for the perturbation that has the minimal $\ell_p$ norm, under the constraint that $L(f_\theta(x + \tau), y)$ is bigger than a given bound $c$:

$$\arg\min_{\tau} \|\tau\|_p \text{ s.t. } L(f_\theta(x + \tau), y) \geq c.$$  

(4)

The value of $c$ is typically chosen depending on the loss function $L$. For example, if $L$ is the 0/1 loss, any $c > 0$ is acceptable. Problem (4) has been tackled by Carlini and Wagner (2017), leading to the strongest method known so far. (Denoted C&W attack in the rest of the paper.) It aims at solving the following Lagrangian relaxation of Problem (4):

$$\min_{\tau} \|\tau\|_p + \lambda \times g(x + \tau),$$

(5)

where $g(x + \tau) < 0$ if and only if $L(f_\theta(x + \tau), y) > c$. The authors use a change of variable $\tau = \tanh(w - x)$ to ensure that $-1 \leq \tau \leq 1$, a binary search to optimize the constant $\lambda$, 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 $\ell_2$ attack. Accordingly, for this work, we only consider C&W as an $\ell_2$ attack solving a norm minimization problem.

2.2 Defense mechanisms

Adversarial Training. Adversarial Training (AT) was introduced by Goodfellow et al. (2015) and later improved by Madry et al. (2018) as a first defense mechanism to train robust neural networks. It consists in augmenting training batches with adversarial examples generated during the training procedure. At each training step, the standard training procedure from Equation 1 is replaced with a $\min \max$ objective function to minimize the expected value of maximum (perturbed) loss, as follows:

$$\min_{\theta} \mathbb{E}_{(x,y) \sim D} \max_{\|\tau\|_p \leq \epsilon} L(f_\theta(x + \tau), y).$$

(6)

In the case where $p = \infty$, this technique offers good robustness against $\ell_\infty$ attacks (Athalye et al., 2018). AT can also be performed using other kinds of attacks (including strong $\ell_2$ attacks such as C&W albeit at a much higher computational cost). However, as we will discuss in Section 3, $\ell_\infty$ adversarial training offers poor protection against $\ell_2$ adversarial attacks and vice-versa.

Noise injection mechanisms. Another important technique to design robust models against adversarial attacks is to inject noise in the model. Injecting a noise vector $\eta$ at inference time results in a randomized neural network $f_\theta := f_\theta(x + \eta)$.

In contrast with Adversarial Training, noise injection mechanisms are, in certain cases, provably robust against adversarial examples as discussed by Pinot et al. (2019); Cohen et al. (2019). Empirical results have also demonstrated their efficiency against $\ell_2$ adversarial attacks (Rakin et al., 2018). These works focus however on Gaussian and Laplace distributions a.k.a. generalized Gaussian of order 2, and 1 respectively. As the limit of a
generalized Gaussian density (Dytso et al., 2018) when \( p \to \infty \) is a Uniform distribution, we also investigate the injection of uniform noise to defend against \( \ell_\infty \) attacks.

3 No Free Lunch for adversarial defenses

3.1 Theoretical analysis

Let us consider a classifier \( f_{\ell_\infty} \) equipped with an ideal defense mechanism against adversarial examples bounded with an \( \ell_\infty \) norm of value \( \epsilon_\infty \). It guarantees that for any input-output pair \((x, y) \sim D\) and for any perturbation \( \tau \) such that \( \|\tau\|_\infty \leq \epsilon_\infty \), \( f_{\ell_\infty} \) is not misled by the perturbation (i.e. \( f_{\ell_\infty}(x + \tau) = f_{\ell_\infty}(x) \)). We now focus our study on the performance of this classifier against adversarial examples bounded with an \( \ell_2 \) norm of value \( \epsilon_2 \).

Using Figure 1(a), we observe that any \( \ell_2 \) adversarial example that is also in the \( \ell_\infty \) ball, is guaranteed to be protected by the \( \ell_\infty \) defense mechanism of \( f_{\ell_\infty} \), but not if it is outside the \( \ell_\infty \) ball. To characterize the probability that an \( \ell_2 \) perturbation is guaranteed to be protected by an \( \ell_\infty \) defense mechanism in the general case (i.e. any dimension \( d \)), we measure the ratio between the volume of intersection of the \( \ell_\infty \) ball of radius \( \epsilon_\infty \) and the \( \ell_2 \) ball of radius \( \epsilon_2 \). 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 \( \ell_\infty \)-bounded adversarial examples, is unlikely to be efficient against \( \ell_2 \) attacks.

**Theorem 1** (Probability of the intersection goes to 0). Let \( B_{2,d} \) and \( B_{\infty,d} \) be two \( d \) dimensional balls, respectively for \( \ell_2 \) norm and \( \ell_\infty \) norm. If for all \( d \), one constrains \( B_{2,d} \) and \( B_{\infty,d} \) to have the same volume, then

\[
\frac{\text{Vol}(B_{2,d} \cap B_{\infty,d})}{\text{Vol}(B_{\infty,d})} \to 0 \text{ when } d \to \infty.
\]

**Proof.** Without loss of generality, let us fix the radius of the \( \ell_\infty \) ball to 1 (denoted \( B_{\infty,d}(1) \)). One can show that for all \( d \), \( \text{Vol}(B_{2,d}(r_2(d))) = \text{Vol}(B_{\infty,d}(1)) \). Where \( r_2(d) = \frac{\sqrt{d}}{\Gamma(d/2) + 1} \), \( \Gamma(d) \) is the gamma function, and \( B_{2,d}(r_2(d)) \) is the \( \ell_2 \) ball of radius \( r_2(d) \). Then, thanks to Stirling’s formula, \( r_2(d) \approx \sqrt{\frac{2}{\pi}}d^{1/2} \). Finally, if we denote \( U_S \), the uniform distribution on \( S \), by using Hoeffding inequality between Eq. (9) and (10), we get:

\[
\frac{\text{Vol}(B_{2,d}(r_2(d)) \cap B_{\infty,d}(1))}{\text{Vol}(B_{\infty,d}(1))} \leq \exp\left\{-d^{-1} \left( r_2^2(d) - d \mathbb{E}[x_1^2] \right)^2 \right\} \leq \exp\left\{-d^{-1} \left( \frac{2}{\pi \epsilon} - \frac{2}{3} \right) d + o(d) \right\}.
\]

Then the ratio between the volume of intersection of the ball and the volume of the ball converges towards 0. \( \Box \)

Theorem 1 states that, when \( d \) is large enough, \( \ell_2 \) bounded perturbations have a null probability of being also in the \( \ell_\infty \) 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 \( \ell_\infty \) adversarial examples is not guaranteed to offer any protection against \( \ell_2 \) attacks, and vice-versa\(^1\).

Remark that this result defeats the 2-dimensional intuition: if we consider a 2 dimensional problem setting, the \( \ell_\infty \) and the \( \ell_2 \) balls have an important overlap (as illustrated in Figure 1(a)) and the probability of sampling in 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 in the intersection of the \( \ell_\infty \) and \( \ell_2 \) balls is close to zero for any realistic image setting. In large dimensions, the volume of the corner of the \( \ell_\infty \) ball is much bigger than it appears in Figure 1(a).

| Dataset   | \( d \) | Inter. (in log, 10) |
|-----------|--------|---------------------|
| –         | 2      | -0.009              |
| MNIST     | 784    | -144                |
| CIFAR     | 3072   | -578                |
| ImageNet  | 150528 | -28946              |

Table 1: Bounds of Theorem 1 on the volume of intersection of \( \ell_2 \) and \( \ell_\infty \) balls at equal volume for typical image classification datasets. When \( d = 2 \), the bound is \( 10^{-0.909} \sim 0.98 \).

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 \( \ell_\infty \) adversarial examples has a null probability of protecting against \( \ell_2 \)-bounded adversarial attacks and vice-versa. 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 w.r.t the result in Theorem 1. As we discuss in Sec. 2, there are two distinctive attack settings: loss maximization (PGD) and perturbation minimization (C&W). We analyse the first setting in details and conduct a second series of experiments to demonstrate that the results are similar if we consider the second setting.

**Adversarial training vs. loss maximization attacks**

To demonstrate that \( \ell_\infty \) adversarial training is not robust against PGD-\( \ell_2 \) attacks, we measure the number of \( \ell_2 \) adversarial examples generated with PGD-\( \ell_2 \), lying outside the \( \ell_\infty \) ball. (Note that we consider all examples, not just the ones that successfully fool the classifier). To do so, we use the same experimental setting as in Section 4 with \( \epsilon_\infty \) and \( \epsilon_2 \) such that the volumes of the two balls are

\(^1\)Th. 1 can easily be extended to any two balls with different norms. For clarity, we restrict to the case of \( \ell_\infty \) and \( \ell_2 \) norms.
equal. Additionally, we also measure the average $\ell_\infty$ and $\ell_2$ norms of these adversarial examples, to understand more precisely the impact of adversarial training, and we report the accuracy, which reflects the number of adversarial examples that successfully fooled the classifier (cf. Table 2 (top)). The same experiment is conducted for $\ell_2$ adversarial training against PGD-$\ell_\infty$ and the results are presented in Table 2 (bottom). All experiments in this section are conducted on CIFAR-10, and the experimental setting is fully detailed in Section 4.4.

![Figure 1](https://via.placeholder.com/150)

Figure 1: Left: 2D representation of the $\ell_\infty$ and $\ell_2$ balls of respective radius $\epsilon$ and $\epsilon'$. Middle: a classifier trained with $\ell_\infty$ adversarial perturbations (materialized by the red line) remains vulnerable to $\ell_2$ attacks. Right: a classifier trained with $\ell_2$ adversarial perturbations (materialized by the blue line) remains vulnerable to $\ell_\infty$ adversarial examples.

| PGD-$\ell_2$ vs. → | Unprotected | AT-$\ell_\infty$ |
|---------------------|-------------|----------------|
| Examples inside $\ell_2$ ball | 100% | 100% |
| Average $\ell_2$ norm | 0.83 | 0.83 |

| PGD-$\ell_\infty$ vs. → | Unprotected | AT-$\ell_2$ |
|-------------------------|-------------|------------|
| Examples inside $\ell_2$ ball | 100% | 100% |
| Average $\ell_2$ norm | 1.4 | 1.64 |

| Accuracy under attack | 0.00 | 0.37 |

Table 2: (Top) number of PGD-$\ell_2$ adversarial examples inside the $\ell_\infty$ and inside the $\ell_2$ ball, without and with $\ell_\infty$ adversarial training. (Bottom) number of PGD-$\ell_\infty$ adversarial examples inside the $\ell_\infty$ and inside the $\ell_2$ ball, without and with $\ell_2$ adversarial training. On CIFAR-10 ($d = 3072$).

The results are unambiguous: none of the adversarial examples generated with PGD-$\ell_2$ are inside the $\ell_\infty$ ball (and thus in the intersection of the two balls). As a consequence, we cannot expect adversarial training $\ell_\infty$ to offer any guaranteed protection against $\ell_2$ adversarial examples. We illustrate this phenomenon using Figure 1 (b): notice that the $\ell_2$ adversarial example represented in this figure cannot be protected using $\ell_\infty$ adversarial training which is only designed to push the decision boundary (red line) outside of the $\ell_\infty$ ball (square), but not outside of the $\ell_2$ ball (circle). Our results demonstrate that all PGD-$\ell_2$ examples are already in this upper area (outside the intersection), before $\ell_\infty$ adversarial training. Therefore $\ell_\infty$ adversarial training is unnecessary.

The second experiment naturally demonstrates a similar behaviour. We first observe that adversarial examples generated with PGD-$\ell_\infty$ lying outside the $\ell_2$ ball cannot be eliminated using $\ell_2$ adversarial training (as illustrated in Figure 1 (c)). However, Table 2 shows that all examples are already outside the $\ell_2$ ball, clustered around the corner of the $\ell_\infty$ ball (average distance is 1.64 compared to $0.031 \times \sqrt{3072} = 1.71$ for the corner). Therefore, any defense method (including $\ell_2$ adversarial training) that would eliminate only adversarial examples inside the $\ell_2$ balls, cannot be efficient against $\ell_\infty$ adversarial examples.

The comparison of accuracy under PGD-$\ell_2$ attack of a classifier defended by either $\ell_\infty$ or $\ell_2$ adversarial training corroborate our analysis. In fact, when defended with AT-$\ell_\infty$, the accuracy of the classifier under attack is 0.37, while the AT-$\ell_2$ defends the classifier up to 0.52 i.e. 40% better. Similarly, a classifier defended with AT-$\ell_\infty$ with an accuracy under PGD-$\ell_\infty$ attack of 0.43 performs 16% better than the one defended with AT-$\ell_2$ which obtains 0.37 accuracy under attack. These results keep confirming our claim: $\ell_2$-based defenses are inadequate to defend against $\ell_\infty$ attacks, and vice-versa.

| Accuracy under attack | 0.00 | 0.37 |

Table 3: This table shows the amount of adversarial examples inside the $\ell_\infty$ ball and inside the $\ell_2$ ball but outside the $\ell_\infty$ ball. We can observe a clear shift between a baseline model (no defense) and a model trained with Adversarial Training PGD $\ell_\infty$ attacked with C&W attack (Carlini and Wagner, 2017).

**Adversarial training vs. perturbation minimization attacks.** We now study the performances of an $\ell_2$ perturbation maximization attack (C&W) with and without AT-$\ell_\infty$. It allows us to understand in which area C&W discovers adversarial examples and the impact of AT-$\ell_\infty$. The results are reported in Table 3. First, when the classifier is undefended, we observe that 70% of adversarial
examples lie inside the intersection of the two balls. This phenomenon is due to the fact that C&W minimizes the $\ell_2$ norm of the perturbation. Therefore without AT, the attack is able to discover adversarial examples that are very close to the original image, where the $\ell_\infty$ and the $\ell_2$ balls overlap. When the model is trained with AT-$\ell_\infty$, we observe a clear shift: 71% of the examples are now outside the $\ell_\infty$ ball, but still inside the $\ell_2$ ball, as illustrated in Figure 1 (b). This means that C&W attack still minimizes the $\ell_2$ norm of the perturbation while updating its search space to ignore the examples in the $\ell_\infty$ ball. Since C&W was always able to discover adversarial examples in this area, AT-$\ell_\infty$ offers no extra benefit in terms of robustness (0% Accuracy). Together, these results and Theorem 1 confirm that $\ell_\infty$-based defenses are vulnerable to $\ell_2$-based perturbation minimization attacks.

4 Building Defenses against Multiple Adversarial Attacks

So far, we have shown that adversarial defenses are able to protect only against the norm they have been trained on. In order to solve this problem, we propose 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. In this section we evaluate several of these defense strategies, and compare them against state-of-the-art attacks using a solid experimental setting (the detailed description of the experimental setting is described in Section 4.4).

4.1 MAT – Mixed Adversarial Training

Earlier results have shown that AT-$\ell_p$ improves the robustness against corresponding $\ell_p$-bounded adversarial examples, and the experiments we present in this section corroborate this observation (See Table 4, column: AT). Building on this observation, it is natural to examine the efficiency of Mixed Adversarial Training (MAT) against mixed $\ell_\infty$ and $\ell_2$ attacks. MAT is a variation of AT that uses both $\ell_\infty$-bounded adversarial examples and $\ell_2$-bounded adversarial examples as training examples.

As discussed by Tramèr and Boneh (2019), 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 $\ell_\infty$ and $\ell_2$, and to use it as a training example, as described in Equation 12:

$$\min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}} \left[ \frac{1}{2\pi \sigma_1^2} \exp \left\{ -\frac{x^2}{2\sigma_1^2} \right\} \mathbb{1}_{\left\{ \|x\|_\infty \leq \sigma_2 \right\}} + \frac{1}{8\pi \sigma_2^2} \exp \left\{ -\frac{x^2}{8\sigma_2^2} \right\} \mathbb{1}_{\left\{ \|x\|_2 \leq \sigma_2 \right\}} \right].$$

(12)

It can be written as:

$$\min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}} \left[ \frac{1}{2\pi \sigma_1^2} \exp \left\{ -\frac{x^2}{2\sigma_1^2} \right\} \mathbb{1}_{\left\{ \|x\|_\infty \leq \sigma_2 \right\}} + \frac{1}{4\pi \sigma_2^2} \exp \left\{ -\frac{x^2}{4\sigma_2^2} \right\} \mathbb{1}_{\left\{ \|x\|_2 \leq \sigma_2 \right\}} \right].$$

(13)

An alternative strategy is to systematically train the model with the most damaging adversarial example ($\ell_\infty$ or $\ell_2$). As described in Equation 13:

$$\min_{\theta} \mathbb{E}_{(x,y) \sim \mathcal{D}} \left[ \frac{1}{2\pi \sigma_1^2} \exp \left\{ -\frac{x^2}{2\sigma_1^2} \right\} \mathbb{1}_{\left\{ \|x\|_\infty \leq \sigma_2 \right\}} + \frac{1}{4\pi \sigma_2^2} \exp \left\{ -\frac{x^2}{4\sigma_2^2} \right\} \mathbb{1}_{\left\{ \|x\|_2 \leq \sigma_2 \right\}} \right].$$

(14)

The accuracy of MAT-Rand and MAT-Max are reported in Table 4 (Column: MAT). As expected, we observe that MAT-Rand and MAT-Max offer better robustness both against PGD-$\ell_2$ and PGD-$\ell_\infty$ 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, AT is very weak against perturbations minimization attacks such as C&W, and MAT is no better against such attacks. This weakness makes MAT of little practical use.

4.2 MNI – Multiple Noise Injection

Another important technique to defend against adversarial examples is to use Noise Injection (NI). Pinot et al. (2019) demonstrated that injecting noise in the network can give provable defense against adversarial examples. Furthermore, we found that NI offers better protection than AT against perturbation minimization attacks such as C&W, thus, they are good candidates to obtain models robust to multiple attacks. In this work, besides the generalized Gaussian noises, already investigated in previous works, we evaluate the efficiency of uniform distributions which are generalized Gaussian of order $\infty$. As shown in Table 4 (Columns: NI), noise injection from this distribution gives better results than Gaussian noise injection against all the attacks except PGD-$\ell_\infty$.

To obtain the best out of both noises, we propose to combine them (MNI) either by convolution (Conv) or by mixture (Mix). Hence, the final noise vector comes from one of the following probability density functions:

$$\frac{1}{\sqrt{2\pi \sigma_1^2}} \exp \left\{ -\frac{x^2}{2\sigma_1^2} \right\} \mathbb{1}_{\left\{ \|x\|_\infty \leq \sigma_2 \right\}} + \frac{1}{\sqrt{8\pi \sigma_2^2}} \exp \left\{ -\frac{x^2}{8\sigma_2^2} \right\} \mathbb{1}_{\left\{ \|x\|_2 \leq \sigma_2 \right\}}.$$

(15)

Following the literature (Pinot et al., 2019), we choose $\sigma_1 = 0.25$. Accordingly, we take $\sigma_2 = 0.2$. The results are presented in Table 4 (Column: MNI). We found that MNI offers comparable results against the experimental setting in (Pinot et al., 2019), but does not improve over NI with a uniform distribution.

4.3 RAT – Randomized Adversarial Training

We now examine the performance of Randomized Adversarial Training (RAT) which mixes Adversarial Training with Noise Injection. We consider the two symmetric settings: RAT-$\ell_\infty$ and a noise from a normal distribution, as well as RAT-$\ell_2$ and a noise from a uniform distribution. The corresponding loss function is defined as follows:

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

(16)

where $f_\theta$ is a randomized neural network with noise injection as described in Section 2.2.
4.4 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 Athalye et al. (2017) and later used by Athalye et al. (2018); Carlini et al. (2019) to identify flawed defense mechanisms.

To attack the models, we use state-of-the-art algorithms PGD and C&W (see Section 2). We run PGD with 20 iterations to generate adversarial examples and with 10 iterations when it is used for adversarial training. We run C&W with 60 iterations to generate adversarial examples. For bounded attacks, the maximum $\ell_\infty$ bound is fixed to 0.031 and the maximum $\ell_2$ bound is fixed to 0.83. As discussed in Section 2, we chose these values so that the $\ell_\infty$ and the $\ell_2$ 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 Zagoruyko and Komodakis (2016). Training time varies from 1 day (AT) to 2 days (MAT) on 4 GPUs-V100 servers.

5 Related Work

Adversarial attacks have been an active topic in the machine learning community since their discovery (Globerson and Roweis, 2006; Biggio et al., 2013; Szegedy et al., 2014). Many attacks have been developed. Most of them solve a loss maximization problem with either $\ell_\infty$ (Goodfellow et al., 2015; Kurakin et al., 2016; Madry et al., 2017; Carlini and Wagner, 2017; Kurakin et al., 2018; Madry et al., 2018) or $\ell_1$ (Tramer and Boneh, 2019) or $\ell_0$ (Papernot et al., 2016) surrogate norms.

Defending against adversarial examples is a challenging problem since the number of layers makes it difficult to understand the geometry of the decision boundary. Despite empirically proven efficient, Adversarial training (Goodfellow et al., 2015) gives no formal defense guarantees. Besides this line of work, randomization and smoothing (Xie et al., 2018; Lecuyer et al., 2018; Pinot et al., 2019; Cohen et al., 2019) have gained popularity since they provide guarantees, but so far, the efficiency of these methods remains limited against $\ell_\infty$-based attacks.

An open question so far is to build an efficient defense against multiple norms. Concurrently to our work, Tramer and Boneh (2019) proposed to tackle this issue by mixing randomized training with attacks for different norms to defend against multiple perturbations. Then, Salman et al. (2019) proposed to mix adversarial training with randomized smoothing to have better certificates against adversarial attacks. These methods are closely related respectively to MAT and RAT. Aside from these similarities, we propose a new geometric point of view for robustness against multiple perturbations, that is backed up theoretically and experimentally. We also conduct a rigorous and full comparison of RAT and MAT as defenses against adversarial attacks. Finally, we propose MNI, that adds mixture of noise to our network and gets promising results. To the best of our knowledge, this is the first work that covers mixtures and convolution of noises with different natures.

6 Conclusion

In this paper, we tackle the problem of protecting neural networks against multiple attacks crafted from different norms. First, we demonstrate that existing defense mechanisms can only protect against one type of attacks. Then we consider a variety of strategies to mix defense mechanisms and to build models that are robust against multiple adversarial attacks. We show that Randomized Adversarial Training offers the best global performance.

| Baseline | AT | MAT | NI | MNI | RAT-\(\ell_\infty\) | RAT-\(\ell_2\) |
|----------|----|-----|----|-----|-------------------|------------------|
| Natural examples | - | $\ell_\infty$ | $\ell_2$ | Max | Rand | $\ell_0$ | Mix | Conv | Baseline | AT | MAT | NI | MNI | RAT-\(\ell_\infty\) | RAT-\(\ell_2\) |
| PGD-\(\ell_\infty\) 20 | 0.94 | 0.85 | 0.85 | 0.80 | 0.80 | 0.79 | 0.87 | 0.84 | 0.79 | 0.74 | 0.85 | 0.79 | 0.87 |
| PGD-\(\ell_2\) 20 | 0.00 | 0.37 | 0.52 | 0.50 | 0.55 | 0.34 | 0.36 | 0.33 | 0.32 | 0.43 | 0.39 | 0.34 | 0.37 |
| C&W-\(\ell_2\) 60 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.33 | 0.53 | 0.41 | 0.32 | 0.30 | 0.41 | 0.33 | 0.34 |

Table 4: This table shows a comprehensive list of results consisting of the accuracy of several defense mechanisms against $\ell_2$ and $\ell_\infty$ attacks. This table main objective is to compare the overall performance of ‘single’ norm defense mechanisms (AT and NI presented in the Sec. 2.2) against mixed norms defense mechanisms (MNI, MAT & RAT mixed defenses presented in Sec. 4). The red values present all accuracy below 30% which shows that all defense mechanisms have ‘weaknesses’ with the exception of RAT.
References

Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. arXiv preprint arXiv:1707.07397, 2017.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In Jennifer Dy and Andreas 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.

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio 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.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In 2017 IEEE Symposium on Security and Privacy (SP), pages 39–57. IEEE, 2017.

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

Jeremy M. Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing. CoRR, abs/1902.02918, 2019.

Alex Dytso, Ronit Bustin, H. Vincent Poor, and Shlomo Shamai. Analytical properties of generalized gaussian distributions. Journal of Statistical Distributions and Applications, 5, 12 2018.

Amir Globerson and Sam Roweis. Nightmare at test time: robust learning by feature deletion. In Proceedings of the 23rd international conference on Machine learning, pages 353–360, 2006.

Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In International Conference on Learning Representations, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.

Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdelrahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Brian Kingsbury, et al. Deep neural networks for acoustic modeling in speech recognition. IEEE Signal processing magazine, 29, 2012.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. arXiv preprint arXiv:1607.02533, 2016.

M. Lecuyer, V. Atlidakis, 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.

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

Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram 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.

Rafael Pinot, Laurent Meunier, Alexandre Araujo, Hisashi Kashima, Florian Yger, Cedric Gouy-Pailler, and Jamal Atif. Theoretical evidence for adversarial robustness through randomization. In Advances in Neural Information Processing Systems 32, pages 11838–11848, 2019.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAi, 2018.

Adnan Siraj Rakin, Zhezhi He, and Deliang Fan. Parametric noise injection: Trainable randomness to improve deep neural network robustness against adversarial attack. CoRR, abs/1811.09310, 2018.

Hadi Salman, Jerry Li, Ilya Razenshteyn, Pengchuan Zhang, Huan Zhang, Sebastien Bubeck, and Greg Yang. Provably robust deep learning via adversarially trained smoothed classifiers. In Advances in Neural Information Processing Systems, pages 11289–11300, 2019.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In International Conference on Learning Representations, 2014.

Florian Tramèr and Dan Boneh. Adversarial training and robustness for multiple perturbations. arXiv preprint arXiv:1904.13000, 2019.

Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. In International Conference on Learning Representations, 2018.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. arXiv preprint arXiv:1605.07146, 2016.