DEGRADATION ATTACKS ON CERTIFIABLY ROBUST NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Certifiably robust neural networks employ provable run-time defenses against adversarial examples by checking if the model is locally robust at the input under evaluation. We show through examples and experiments that these defenses are inherently over-cautious. Specifically, they flag inputs for which local robustness checks fail, but yet that are not adversarial; i.e., they are classified consistently with all *valid* inputs within a distance of ϵ. As a result, while a norm-bounded adversary cannot change the classification of an input, it *can* use norm-bounded changes to degrade the utility of certifiably robust networks by forcing them to reject otherwise correctly classifiable inputs. We empirically demonstrate the efficacy of such attacks against state-of-the-art certifiable defenses.

1 INTRODUCTION

An adversarial example for a neural classifier is the result of applying small modifications to a correctly classified *valid* input such that the modified input is classified incorrectly. For neural classifiers trained in a standard manner, it has been shown that adversarial examples are rampant (Szegedy et al., 2014; Carlini & Wagner, 2017). While training models in an adversarially-aware manner helps mitigate the issue to an extent, it does not solve the problem, as evidenced by the large gaps between the clean accuracy and verified robust accuracy[1] of models.

Certifiably robust classifiers (Weng et al., 2018; Wong et al., 2018; Cohen et al., 2019; Leino et al., 2021) offer the most rigorous solution to the problem and provably protect models against adversarial attacks. These classifiers are constructed by composing a standard classifier with a certified run-time defense. The defenses aim to detect adversarial examples during model evaluation, by checking if the model is ϵ-locally robust at the evaluated input. If the check fails, the input is flagged as (potentially) adversarial and the model rejects the input. However, a rejection is not free of cost. Every time the model rejects an input, a user of the model has to resort to some other process other than their presumed first choice (the defended model) to make a prediction, reducing model utility.

We show that existing certified run-time defenses are overly cautious and susceptible to erroneously flagging non-adversarial inputs as adversarial. This over-cautiousness is inherent in the design of such defenses, and it is manifested even when the ϵ-local robustness check is exact. Not only does this lead to a degradation in model utility because of unnecessary rejections, but it also exposes models to a new line of attack that we refer to as *degradation attacks*. We develop new attacks that are aimed at causing certifiably robust classifiers to reject inputs frequently, and we show that this is a significant problem in practice. For state-of-the-art certifiably robust models, such as GloRo Nets (Leino et al., 2021) and randomized smoothed models (Cohen et al., 2019), our attacks succeed on as many as 56% of the inputs where the model is already known to be robust. This is particularly distressing given the already considerable computational costs of training and defending certifiably robust models.

As a concrete scenario, consider an autonomous driving system that uses a neural classifier for labeling road signs. It is unsafe for the classifier to misclassify adversarially perturbed road signs, and we want to prevent this at all costs. One simple strategy is for the classifier to always reject inputs and hand over decision-making to the human driver. This is perfectly safe behavior, but this model has zero utility. Ideally, we want the model to hand over control (i.e., the run-time defense should raise a flag) only when the perturbed example is actually going to cause misclassification. Our result implies that an adversary can cause the model to hand over control to the human driver even when a perturbed input would not have

[1] % of test inputs where the model is accurate and also locally robust

been misclassified. Although this does not happen for all the inputs (as in our simple example), it happens often enough that the adversary can cause a drastic and unnecessary reduction in model utility.

Degradation attacks are successful because certified run-time defenses do not account for the data manifold. An adversarial example is obtained by applying modifications to *valid* inputs. We formalize valid inputs by means of a set M which is the support set of the underlying input distribution that characterizes the learning problem. Although the exact description of M is unknown, we know that both the training and test data used are in M. In order for a model to be free of adversarial examples, it only needs to be locally robust at every point in M. Certified defenses, however, try to enforce local robustness at all evaluated inputs, irrespective of whether the input belongs to M. An incorrectly flagged non-adversarial input is then an input at which the local robustness check fails but if the model were not to reject the input, it would be classified consistently with all valid inputs within an ϵ distance. While a norm-bounded adversary cannot change the classification of an input in the presence of a certified run-time defense, it can apply norm-bounded modifications to existing valid inputs (i.e., inputs in M) and force the model to reject otherwise correctly classifiable inputs, thereby degrading the model's utility.

As we discuss in Section 4, defenders against degradation attacks have two options: they can use existing methods for constructing certifiably robust models but with double the radius that the adversary is allowed for perturbations, or they can develop new defense techniques that account for the fact that models only need to be locally robust at points in M. We evaluate the ramifications of the former option and leave the latter for future work.

To summarize, the main contributions of our work are as follows: (1) we describe new attacks, referred to as *degradation attacks*, that can force certifiably robust models (with either deterministic or stochastic defense mechanisms) to frequently and unnecessarily reject inputs; (2) we empirically demonstrate the severity of the problem for models using state-of-the-art certified defense mechanisms like GloRo Nets and Randomized Smoothing; (3) we make explicit the set, M, from which *valid* inputs are drawn, and this helps us explain that certified run-time defenses, based on checking ϵ-local robustness, are susceptible to degradation attacks, as they enforce local robustness at all inputs and not just inputs in M; and (4) we discuss two possible defenses against degradation attacks, and evaluate the one based on doubling the radius enforced by certifiably robust models.

The rest of this paper is organized as follows. In Section 2, we revisit definitions of adversarial robustness and sketch a general approach for certified defenses. In Section 3, we demonstrate how a certified defense can unnecessarily reject inputs; we present degradation attack algorithms to this end, and evaluate their efficacy. In Section 4, we sketch two defenses against degradation attacks, and concretely evaluate one of these proposals. In Section 5, we describe the related work. Finally, we conclude in Section 6.

2 Revisiting Definitions of Adversarial Robustness

A neural classifier $f \in \mathbb{R}^d \rightarrow L$ is a function from a d-dimensional real vector space, $\mathbb{R}^d$, to a finite set of labels, L. The inputs to the classifier are drawn from some distribution with support set $M \subseteq \mathbb{R}^d$. In other words, the training and test sets are comprised of elements from M and their corresponding labels. Though the description of M is not available to us, it plays a key role in formalizing the notion of adversarial examples.

An adversarial example for a neural classifier is the result of applying small modifications to a correctly classified *valid* input such that the modified input is classified incorrectly. Formally, an input is valid if it belongs to M. Definition 1 below, proposed by Szegedy et al. (2014), formalizes the notion of a targeted adversarial example.

Definition 1 (Targeted Adversarial Example) *Given a neural classifier $f \in \mathbb{R}^d \rightarrow L$ and an input $x \in \mathbb{R}^d$, an input $x' := x + r$ is an adversarial example with respect to a target label $l \in L$, ℓ_p metric, and a fixed constant $\epsilon \in \mathbb{R}^d$ if the solution r to the following optimization problem is such that $||r||_p \leq \epsilon$:*

$$\textit{Minimize } ||r||_p \textit{ subject to } f(x+r) = l \textit{ and } x+r \in [0,1]^d$$

One need not assume that the inputs are scaled to be in $[0,1]^d$, and hence we can drop the second optimization constraint. *Untargeted adversarial examples*, or simply *adversarial examples*, are defined similarly except that the constraint $f(x+r) = l$ is replaced with $f(x+r) \neq f(x)$.

Algorithm 2.1: Prediction with a certified run-time defense

Inputs: A model $f \in \mathbb{R}^d \rightarrow L$, an input $x' \in \mathbb{R}^d$, an attack bound $\epsilon \in \mathbb{R}$, and a distance metric ℓ_p
Output: A prediction $l \in L \cup \{\perp\}$

```
1 PredictWithDefense(f, x', ε, ℓp):
2     if f is (ε,ℓp)−locally robust at x' then
3         return f(x')
4     else
5         return ⊥
```

As stated informally in (Carlini & Wagner, 2017; Szegedy et al., 2014) x needs to be a *valid* input, i.e., $x \in M$, for x' to be an adversarial example. Making this requirement explicit and formal (via set M) is a key step in the design of our degradation attacks.

A classifier is protected from adversarial examples with respect to a valid input x if it is locally robust at x. As stated in Definition 2, a classifier is locally robust at x, if its prediction does not change in an ϵ-ball centered at x.

Definition 2 (Local Robustness) *A neural classifier $f \in \mathbb{R}^d \rightarrow L$ is (ϵ,ℓ_p)-locally robust at $x \in \mathbb{R}^d$ if,*

$$\forall x' \in \mathbb{R}^d.\ ||x - x'||_p \leq \epsilon \Longrightarrow f(x) = f(x')$$

The formal notions of validity and local robustness enable us to make the following key observation.

Observation 3 *A classifier is protected from **all** adversarial examples if it is locally robust at **all** points in M.*

Certified run-time defenses. Classifiers, when composed with certified run-time defenses, are guaranteed protection from adversarial examples. These defenses, briefly surveyed in Section 5, detect adversarial examples during model evaluation (or run-time) by checking if the model is locally robust at the input being evaluated. Algorithm 2.1 describes the manner in which certified run-time defenses are deployed for protection.

The common assumption is that an adversary has white-box access to f as well as to valid inputs (denoted x), and that the adversary may or may not succeed in building adversarial examples (denoted x') that are ϵ-close to x under the ℓ_p metric. Given input x', a certified run-time defense helps decide whether x' is an adversarial example or not. As described by `PredictWithDefense` (lines 1-5), the defense mechanism is first used to check if the model f is locally robust at x' (line 2). If the check passes, then x' is guaranteed to not be an adversarial example and the model prediction at x' is returned as the label (line 3). However, if the check fails, then x' maybe an adversarial example, though it not guaranteed to be one. The model conservatively rejects the input, indicated here as returning the $\perp$ label (line 5).

3 Degradation Attacks

In this section, we describe our degradation attack algorithms and demonstrate their efficacy empirically. First, in Section 3.1, by means of an example, we demonstrate how certified defenses can lead to unnecessary rejections. Next, in Section 3.2, we describe our degradation attack algorithms against deterministic and stochastic certified defenses. Finally, in Section 3.3, we evaluate our attacks empirically.

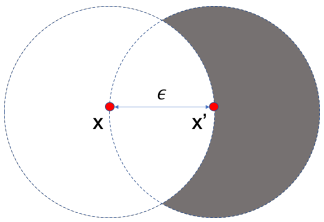

Figure 1: An erroneously flagged non-adversarial example

3.1 An example of a certified defense failure

Certified run-time defenses can erroneously flag a non-adversarial input as adversarial because they enforce local robustness at all inputs and not just inputs in M. Consider the example in Figure 1.

Algorithm 3.1: Degradation attack algorithm

Inputs: A model $f \in \mathbb{R}^d \rightarrow L$, an input $x \in M$, an attack bound $\epsilon \in \mathbb{R}$, and a distance metric ℓ_p
Output: An attack input $x' \in \mathbb{R}^d$

1 `DegradationAttack(`f , x , ϵ , ℓ_p`):`
2 $\quad x'' :=$ `Attack(`f,x,2ϵ,ℓ_p`)`
3 $\quad x' := x + \epsilon \cdot (x'' - x) / ||x'' - x||_p$
4 $\quad$ **return** x'

Algorithm 3.2: Smoothed projected gradient descent attack (SPGD)

Inputs: A model $f \in \mathbb{R}^d \rightarrow \mathbb{R}^{|L|}$ mapping inputs to logit outputs, a loss function $\mathcal{L}$, an input $x \in M$, an attack bound $\epsilon \in \mathbb{R}$, a distance metric ℓ_p, a step size η, a number of steps N, a number of samples n, and a noise parameter σ
Output: An attack input $x' \in \mathbb{R}^d$

1 `SmoothedPgdAttack(`f , $\mathcal{L}$, x , ϵ , ℓ_p , η , N , n , σ`):`
2 $\quad x' := x$
3 $\quad$ **for** $0 \leq \textit{step} < N$ **do**
4 $\quad\quad \zeta \sim \mathcal{N}(0,\sigma)^{n \times d}$
5 $\quad\quad x' := x' + \nabla_{x'} \mathcal{L}\big(\frac{1}{n} \sum_i f(x' + \zeta_i)\big)$
6 $\quad\quad x' := x + \min\{\epsilon , ||x' - x||_p\} \cdot \frac{x' - x}{||x' - x||_p}$
7 $\quad$ **return** x'

x is a valid input from M. Let us assume that none of the points in the gray region belong to M. A binary classifier $f \in \mathbb{R}^2 \rightarrow \{W,G\}$ assigns label W to all points in the white region (which includes x and x') and label G to all points in the gray region. x and x' are ϵ apart and the radius for both the circles is ϵ. The (ϵ,l_2)-local robustness check at x' will fail since the ϵ-ball centered at x' includes white and gray points. However, x' is not an adversarial input since it shares its label with all the ϵ-close inputs in M. Consequently, the model rejection of x' is unnecessary and causes a degradation in model utility. This happens even if the local check is precise and the bound ϵ is tight

3.2 ALGORITHMS

Algorithm 3.1 describes a simple degradation attack against a deterministic defense. Intuitively, given a valid input x (on which the classifier f is (ϵ,ℓ_p)-locally robust, for some fixed ϵ and distance metric ℓ_p), we want to find an input x' such that $||x - x'||_p \leq \epsilon$ and the classifier f is not (ϵ,ℓ_p)-locally robust at x'. The attack re-purposes existing white-box attack algorithms (like Carlini-Wagner and PGD) that search for adversarial examples. In particular, the attack invokes an off-the-shelf white-box attack algorithm, indicated by `Attack`, to find an adversarial example x'' in a ball of radius 2ϵ centered at x (line 2). Next, x'' is simply projected onto the ϵ-sphere centered at x (line 3). As our empirical results in Section 3.3.1 demonstrate, even this seemingly simple attack can be highly effective in degrading model utility.

Attacking stochastic defenses. Stochastic defenses, such as randomized smoothing, transform a base classifier into a smoothed classifier by convolving the base classifier with a Gaussian distribution. The smoothed classifier does not have an explicit representation, and to evaluate it on a single input, the base classifier needs to evaluated on as many as 100,000 samples (Cohen et al., 2019). Moreover, the lack of explicit function representation precludes the use of off-the-shelf white-box attack algorithms that perform gradient descent in the input space. While one could simply perform a standard white-box attack on the base classifier, we found it was more effective to implicitly attack the smoothed model by taking the average gradient over a set of randomly generated samples at each gradient descent step. This gives us a "smoothed PGD" (SPGD) attack tailored for use against stochastic defenses. Our SPGD procedure is given in Algorithm 3.2.

3.3 EVALUATING ATTACK EFFICACY

Our empirical evaluation is designed to measure the susceptibility of state-of-the-art robust models and certified run-time defenses to utility degradation attacks. For our experiments, we consider GloRo Nets (Leino et al., 2021) and randomized smoothed models (Cohen et al., 2019). These approaches lead to models with the best known verified robust accuracies (VRA), with respect to the ℓ_2 metric, on a variety of popular image classification datasets like MNIST (LeCun et al., 2010), CIFAR-10 (Krizhevsky, 2009), and ImageNet (Deng et al., 2009).

GloRo Nets incorporate global Lipschitz bound computation into the model architecture. The global Lipschitz bound is used to implement the local robustness check at an input under evaluation. Moreover, these local robustness checks are backpropagated over during training to encourage learning of models with a high degree of robustness. In contrast, randomized smoothing transforms a base classifier into a smoothed classifier by convolving the base classifier with a Gaussian distribution. The smoothed classifier typically demonstrates a much higher degree of robustness than the base classifier as long as the base classifier has been trained to classify well under Gaussian noise. The method uses statistical techniques to compute the local (certified) robustness radius R of the smoothed classifier at a given input, so an (ϵ,ℓ_2)-local robustness check simply amounts to checking if the radius R at the evaluated input is greater than ϵ or not.

We conducted two sets of experiments. In our first set of experiments, described in Section 3.3.1, we attack certifiably robust models at inputs in the test set using the approaches described in Section 3.2. This gives a lower bound on the efficacy of degradation attacks. Our second set of experiments, described in Section 3.3.2, measure the susceptibility of models to degradation attacks assuming an all-powerful adversary, i.e., an adversary that succeeds whenever the model is susceptible to a degradation attack at inputs in the test set. This provides an upper bound on the efficacy of degradation attacks.

We train GloRo Nets using the publicly available code[2]. For randomized smoothed models, we use the pre-trained models made available by the authors [3]. The GloRo Nets used in the evaluation are composed of convolution and fully-connected layers. For MNIST ($\epsilon = 0.3$), the model has two convolution layers and two fully-connected layers (2C2F), for MNIST ($\epsilon = 1.58$), the model is 4C3F, and for CIFAR-10 ($\epsilon = 0.141$), the model is 6C2F. For CIFAR-10, the randomized smoothed model uses a 110-layer residual network as the base classifier, and for ImageNet, a ResNet-50 model is used as the base classifier We implemented our attacks in Python, using TensorFlow and PyTorch. All our experiments were run on an NVIDIA TITAN RTX GPU with 24 GB of RAM, and a 4.2GHz Intel Core i7-7700K with 32 GB of RAM.

3.3.1 LOWER BOUNDS ON ATTACK EFFICACY

To measure the efficacy of degradation attacks on a particular model and dataset, we construct two subsets of the test set for the given dataset, that we refer to as test_R and test_A. The former, test_R, is the set of test inputs on which the model is (ϵ,ℓ_p)-locally robust. The latter, test_A, is a subset of test_R such that for each input $x \in \mathit{test}_A$ the model is (ϵ,ℓ_p)-locally robust at x but there exists an input x' in the ϵ-ball centered at x such that the model is not (ϵ,ℓ_p)-locally robust at x'. Since the model is locally robust at x, we know that x and x' share the same label and x' is not an adversarial input (assuming that there exists no other point in M, except x, that is ϵ-close to x'). Yet, the model would reject x' since the local robustness check fails at x'. Thus, test_A represents the set of test inputs where the model is susceptible to degradation attacks. The ratio $|\mathit{test}_A|/|\mathit{test}_R|$ represents the *false positive rate* of the certified run-time defense, i.e., the rate at which the certified defense erroneously causes the model to reject inputs. We measure model susceptibility and attack efficacy in terms of the false positive rate. A value of one indicates that the model is susceptible to a degradation attack on all test inputs where it is locally robust, while a value of zero indicates that model is safe from degradation attacks on all test inputs.

Table 1 presents lower bounds on degradation attack efficacy against GloRo Nets, measured by performing our degradation attack described in Algorithm 3.1. Table 2 presents lower bounds on degradation attack efficacy against Randomized Smoothing, measured by performing our SPGD degradation attack described in Algorithm 3.2.

We see that against both GloRo Nets and Randomized Smoothing our attacks decrease model utility to varying degrees. In some cases the reduction in utility was substantial, with false positive rates at 50% or

[2] `https://github.com/klasleino/gloro`

[3] `https://github.com/locuslab/smoothing`

	baseline VRA (%)	*false positive rate* (%)	*utility reduction* (%)
MNIST (ϵ=0.3)	95.0	6.0	5.7
MNIST (ϵ=1.58)	61.8	56.1	34.7
CIFAR-10 (ϵ=0.141)	60.0	11.4	6.6

Table 1: Lower bounds for false positive rates induced by degradation attacks against GloRo Nets.

	σ	*baseline VRA* (%)	*false positive rate* (%)	*utility reduction* (%)
CIFAR-10 (ϵ=0.5)	0.25	48.0	50.0	24.0
CIFAR-10 (ϵ=0.5)	0.50	41.0	29.3	12.0
ImageNet (ϵ=1.0)	0.50	46.0	17.4	8.0

Table 2: Lower bounds for false positive rates induced by degradation attacks against Randomized Smoothing. Similarly to the evaluation of Cohen et al. (2019) results are obtained on a sample of 100 arbitrary test points, as Randomized Smoothing is costly to evaluate.

	clean accuracy (%)		*ϵ-VRA* (%)		*2ϵ-VRA* (%)		*ϵ -rejection rate* (%)		*2ϵ-rejection rate* (%)		*false positive rate* (%)	
	S	**D**	**S**	**D**	**S**	**D**	**S**	**D**	**S**	**D**	**S**	**D**
MNIST (ϵ=0.3)	99.15	98.99	95.04	96.99	73.98	91.67	4.91	2.79	26.02	8.3	22.20	5.67
MNIST (ϵ=1.58)	98.32	98.99	61.81	61.84	0.36	9.06	38.19	38.11	99.64	90.94	99.42	85.36
CIFAR-10 (ϵ=0.141)	75.17	66.71	60.01	57.88	42.96	48.72	32.18	22.17	54.63	40.13	33.10	23.08

Table 3: Upper bounds for false positive rates on GloRo Nets. **S** indicates models trained for ϵ-local robustness. **D** indicates models trained for 2ϵ-local robustness.

more on MNIST (GloRo Net) and CIFAR-10 (Randomized Smoothing). Overall, false positive rates were higher when the robustness radius was relatively higher; e.g., the false positive rate on MNIST GloRo Nets increases from 6% to 56% when the radius ϵ increased from 0.3 to 1.58. This may be in part because at smaller radii, many points may be likely to be far more robust than required.

Despite the fact that an explicit representation of smoothed models is not available even to a white-box attacker, we find that our SPGD attack was able to successfully degrade the utility of models defended by Randomized Smoothing. On smoothed models we observe that models trained with a larger noise parameter σ are less affected by false positives; however, this comes at the cost of a notable decrease in baseline utility.

3.3.2 Upper bounds on attack efficacy

To compute upper bounds on the efficacy of degradation attacks, i.e., upper bounds on the false positive rates, we construct test_R and test_A as follows. To construct test_R, we simply apply the (ϵ, ℓ_p)-local robustness check at every input in the test set. If the check passes at an input x, it is added to test_R. To construct test_A, we apply a $(2\epsilon,\ell_p)$-local robustness check at every input in test_R. If this stronger check passes at an input x, it means that no matter how the adversary perturbs x within the ϵ-ball centered at x, the model is always (ϵ,ℓ_p)-locally robust at the perturbed input and so cannot be forced to unnecessarily reject the input. However, if the $(2\epsilon,\ell_p)$-local robustness check fails at x, then even though the model is (ϵ,ℓ_p)-locally robust at x, there exists an input x' in the ϵ-ball at x where the model is not (ϵ,ℓ_p)-locally robust and an adversary can force an unnecessary rejection. Therefore, if the stronger check fails at x, we add it to test_A.

Table 3 shows the upper bounds on the false positive rates for GloRo Nets. The rejection rate is the percentage of test inputs where the model was found to be not locally robust. We report VRA and rejection rates at ϵ and 2ϵ. We see that, assuming an all-powerful adversary, GloRo Nets are quite susceptible to degradation attacks. By comparing the upper bounds to the lower bounds obtained in Section 3.3.1, we see that our degradation attacks recover between $1/4$ and $1/2$ of the best possible degradation.

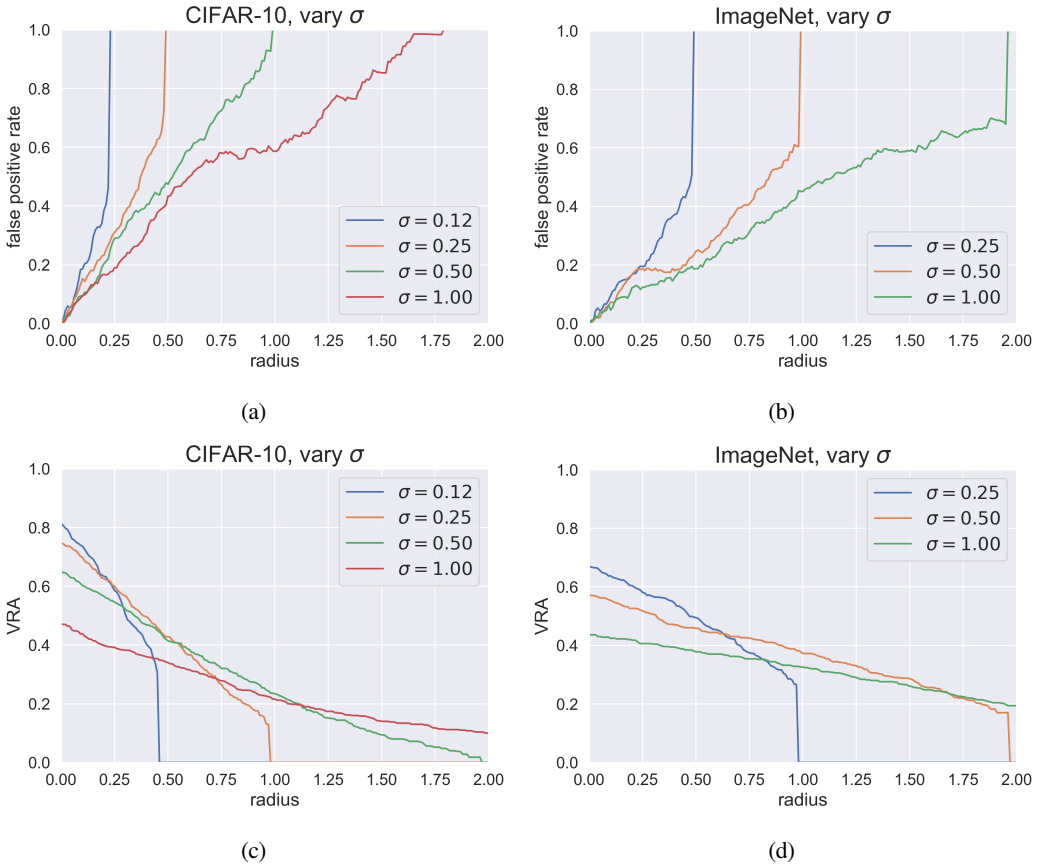

Figure 2: Upper bounds for false positive rates on Randomized Smoothing

Interestingly the susceptibility to degradation attacks does not appear to result from the GloRo Net's overapproximation of local robustness. Leino et al. (2021) report that the Lipschitz bounds on the MNIST models are far tighter than on the CIFAR-10 model, yet the CIFAR-10 model is not more susceptible to degradation. As explained in Section 1, this is because the threat of degradation attacks arises intrinsically from any defense that filters based on local robustness at run-time, *not* because of the overapproximate nature of practical defenses.

Figures 2a and 2b show the false positive rates at different radii (i.e., ϵ values) for each of the randomized smoothed models. The false positive rates are quite high and consistently approach one as ϵ increases. Models that are smoothed using larger σ values unsurprisingly demonstrate more resilience to degradation attacks. However, the greater resilience comes at the cost of a reduced VRA, as shown in the Figures 2c and 2d that plots the VRA of the various models at different radii. The sudden jumps in the false positive rate to one indicate the ϵ value where all points in $test_R$ also end up in $test_A$; i.e., the model is susceptible to degradation attacks at all the points where it is locally robust. A plot for the randomized smoothed models as compared to a table for the GloRo Nets is simply a reflection of the slightly different ways in which these models are evaluated. On each input, a randomized smoothed model outputs a certifiably robust radius along with the prediction, and one then checks if this radius is greater than the desired ϵ or not. In contrast, the desired ϵ is baked in the GloRo Net architecture, and during evaluation, a GloRo Net directly outputs $\perp$ as its prediction if the model is not ϵ-locally robust at the input being evaluated.

4 Defending against degradation attacks

In this section, we sketch two possible ways of defending against degradation attacks, and empirically evaluate the effectiveness of our first proposal.

4.1 DEFENSE VIA RADIUS DOUBLING

A simple strategy to defend against degradation attacks is to enforce local robustness, when training and evaluating certifiably robust models, using twice the radius that the norm-bounded adversary is allowed for adversarial and degradation attacks. This strategy is motivated by the intuition that if a model is 2ϵ robust at a point x in M, then no matter how the adversary perturbs x within the ϵ-ball centered at x, the model is always (ϵ,ℓ_p)-locally robust at the perturbed input and cannot be forced to unnecessarily reject the input.

Training a model for 2ϵ-robustness is not guaranteed to help because of the frequently observed inverse relationship between model accuracy and robustness. The model may be less accurate than the one trained for ϵ-robustness, and may also reject more frequently, even on points in M. However, evaluating the model against 2ϵ is certainly helpful. The 2ϵ-VRA on validation data is a lower bound on the accuracy one can expect on test data, in the face an ϵ-adversary capable of adversarial as well as degradation attacks. This also suggests that, when evaluating certified defenses, one ought to report not just ϵ-VRA, but also 2ϵ-VRA. Evaluating against 2ϵ can also help in configuring the certified run-time defense. In case the rejection rate is too high at a particular 2ϵ, the certified defense can be configured to enforce local robustness at a smaller ϵ value. This reduces vulnerability to degradation attacks at the cost of an increased vulnerability to adversarial attacks.

We evaluate this defense strategy on GloRo Net models using the same architectures and datasets as used for our experiments in Section 3.3. However, we re-train all the models using twice the ϵ values that we want to defend against. Table 3 summarizes the results. Columns labeled **S** show the results for models trained using ϵ, and columns labeled **D** show the results for models trained using 2ϵ. For all the models, we see that training against 2ϵ-robustness helps improve the 2ϵ-VRA and 2ϵ-rejection rates, leading to a reduction in the false positive rates. Moreover, this improvement is achieved without significantly affecting the ϵ-VRA of the models. These results suggest that training against 2ϵ-robustness can reduce the vulnerability to degradation attacks, but the false positive rates still remain concerning.

While smoothed models are not explicitly trained for a specific degree of robustness, and therefore this approach does not directly apply, the noise parameter, σ, does allow one to control the trade-off between robustness and accuracy. Cohen et al. (2019) found that the best VRA is typically achieved when $\sigma \approx \epsilon/2$. Table 2 shows that by increasing σ beyond this point, one can obtain better resistance to degradation. However, as Figures 2c and 2d show, increasing σ also negatively impacts utility.

4.2 DEFENSE VIA AN M-MEMBERSHIP ORACLE

Certified run-time defenses which rely on local robustness checks are susceptible to degradation attacks because these defenses are overly strong; i.e., they try to enforce local robustness at all evaluated inputs, even though they only need to enforce local robustness at inputs in M. To 'fix' these certified defenses, instead of checking local robustness at the given input x', one could instead check if x' shares its label with all the inputs in M that are ϵ-close to x'. If the check passes, then x' is guaranteed to be a non-adversarial input, but, more importantly, if the check fails then x' is *guaranteed* to be an adversarial input.

This fix works because an input x' is an adversarial example if and only if there exists some input $x \in M$ that is ϵ-close to x' and is labeled differently from x'. To implement our proposed fix, we would need an explicit representation for M, or at least an oracle for membership in M. A membership oracle may be used to search for members of M that are ϵ-close to a given x'. While an exhaustive search for all such members is likely infeasible, we may allocate a fixed budget of computational resources to the oracle-based check. If, within the allocated budget, we find an input $x \in M$ such that x is ϵ-close to x' but it has a different label, then we have detected an adversarial example and can reject x'. However, if the check fails to find such an x within the budget, we can fall back to checking local robustness at x'. Practically, one could use an Out-of-Distribution (OOD) detector as an M-membership oracle. Although OOD detectors can be imprecise and susceptible to adversarial attacks themselves (Sehwag et al., 2019; Bitterwolf et al., 2020), some recent developments (Chen et al., 2020) look promising. We leave this for future work.

5 RELATED WORK

The discovery of adversarial examples has led to an intense effort in recent years towards formalizing the problem, as well as designing new attack and defense mechanisms.

Formalizations. The presence of adversarial examples has been popularly formalized as the absence of local robustness in a neural classifier. However, only constant functions are locally robust everywhere, so this definition fails to capture the notion of a classifier free from adversarial examples. Leino et al. (2021) propose the notion of global robustness for classifiers that are allowed to reject inputs. Global robustness requires that the classifier maintain a separation of at least 2ϵ between input regions that are assigned different prediction labels and reject inputs between such regions. Moreover, Yang et al. (2020b) show that in natural image datasets, differently labeled images are indeed separated by values of 2ϵ larger than the perturbation radii used in adversarial example experiments. These observations demonstrate that for real-world datasets, the support set of the input distribution is much smaller than the entire input space and inputs belonging to different classes are separated. A different notion of global robustness is proposed by Ruan et al. (2019). They define a model to be globally robust if it is locally robust at every input in a finite test set. None of these definitions capture the idea that, to be protected from adversarial examples, a model only needs to be locally robust at every input in the support set, M.

Adversarial attacks. Algorithms for constructing adversarial attacks can be divided in to two major categories. White-box attacks assume that they have access to the model internals, i.e., the model architecture and weights. Their primary strategy is to perform gradient descent in the input space so as to find an input that maximizes the loss while satisfying the constraint of staying within the ϵ-ball centered at the original input. Some popular and successful attack algorithms of this nature include the ones proposed by Goodfellow et al. (2015), Carlini & Wagner (2017), and Madry et al. (2018). Black-box attacks only assume query access to the model (Papernot et al., 2017). In other words, their threat model is weaker than the one assumed by white-box attacks. The degradation attacks proposed in this paper use adversarial attack algorithms as a sub-procedure and are agnostic to whether these algorithms are white-box or black-box.

Heuristic and certified defenses. Heuristic defenses against adversarial examples do not provide any guarantees about their defensive capabilities. These include approaches that modify the training objectives, modify the neural classifier post-training, or embed run-time checks to flag adversarial examples during evaluation. The lack of guarantees suggests that these defenses can be broken and it has indeed been demonstrated (Athalye et al., 2018; Tramer et al., 2020) that a number of published defenses are breakable. In this paper, we focus on certified run-time defenses that are deployed during model evaluation and are guaranteed to detect adversarial examples. Such defenses check if the classifier is locally robust at the evaluated input. Guaranteed local robustness checks are implemented using a variety of approaches that include constraint solving and formal methods (Huang et al., 2017; Katz et al., 2017; Gehr et al., 2018; Singh et al., 2019), optimization (Bastani et al., 2016; Dvijotham et al., 2018; Raghunathan et al., 2018; Wong & Kolter, 2018; Tjeng et al., 2019), Lipschitz bounds computation (Weng et al., 2018; Leino et al., 2021), and stochastic smoothing (Lecuyer et al., 2019; Cohen et al., 2019; Yang et al., 2020a). We show that such defenses are overly cautious and often flag even non-adversarial inputs as adversarial.

The two algorithms (PREDICT and CERTIFY) from (Cohen et al., 2019) can both return ABSTAIN when the prediction (and the certified radius) can not be computed for the given input, due to the imprecision of the statistical analysis. This can be potentially exploited by an adversary who can force a certifiably robust classifier to unnecessarily abstain as has been noted in (Cohen et al., 2019). Our attack is different (but similar in spirit) in that it works in cases that CERTIFY does not abstain and is able to reliably compute the prediction and the certified radius R, but this radius may be smaller than the desired ϵ for inputs that are in fact non-adversarial.

6 Conclusion

In this paper, we have showed that certified run-time defenses against adversarial examples, based on local robustness checks, are inherently overcautious and can reduce model utility due to unnecessary input rejections. This is a consequence of these defenses enforcing local robustness at all evaluated inputs even though local robustness only needs to be enforced at valid inputs, i.e., inputs in the support set, M, of the input distribution. An adversary can exploit this over-cautiousness; though they cannot change the classification of an input, they can apply norm-bounded modifications to valid inputs and force the model to reject an otherwise correctly classifiable inputs. We have presented concrete degradation attacks of this nature that can be implemented using off-the-shelf white-box attack algorithms. Moreover, our empirical evaluation demonstrates that even state-of-the-art certifiably robust models, like randomized smoothed models and GloRo Nets, are highly susceptible to utility degradation attacks.

REPRODUCIBILITY STATEMENT

We will make our attack code and experiment scripts available via GitHub. For all our experiments with randomized smoothed models, we directly use the pre-trained models made available by the authors of the paper at `https://github.com/locuslab/smoothing`. For the experiments with GloRo Nets, we trained the models using the publicly available code for training GloRo Nets at `https://github.com/klasleino/gloro`.

ETHICS STATEMENT

Our work sheds light on existing vulnerabilities in state-of-the-art certifiably robust neural classifiers. These degradation attacks could be deployed by malicious entities to degrade the utility of deployed models. However, by putting this knowledge out in the public domain and making practitioners aware of the existence of degradation attacks, we hope that precautions can be taken to protect existing systems. Moreover, it highlights the need to harden future systems against such attacks.

REFERENCES

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pp. 274–283. PMLR, 2018.

Osbert Bastani, Yani Ioannou, Leonidas Lampropoulos, Dimitrios Vytiniotis, Aditya V. Nori, and Antonio Criminisi. Measuring neural net robustness with constraints. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 26212629, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Julian Bitterwolf, Alexander Meinke, and Matthias Hein. Certifiably adversarially robust detection of out-of-distribution data. *Advances in Neural Information Processing Systems*, 33, 2020.

N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57, Los Alamitos, CA, USA, may 2017. IEEE Computer Society. doi: 10.1109/SP.2017.49. URL `https://doi.ieeecomputersociety.org/10.1109/SP.2017.49`.

Jiefeng Chen, Yixuan Li, Xi Wu, Yingyu Liang, and Somesh Jha. Robust out-of-distribution detection for neural networks. *arXiv preprint arXiv:2003.09711*, 2020.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1310–1320. PMLR, 09–15 Jun 2019. URL `https://proceedings.mlr.press/v97/cohen19c.html`.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Krishnamurthy Dvijotham, Robert Stanforth, Sven Gowal, Timothy Mann, and Pushmeet Kohli. A dual approach to scalable verification of deep networks. In *Proceedings of the Thirty-Fourth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-18)*, pp. 162–171, Corvallis, Oregon, 2018. AUAI Press.

Timon Gehr, Matthew Mirman, Dana Drachsler-Cohen, Petar Tsankov, Swarat Chaudhuri, and Martin Vechev. Ai2: Safety and robustness certification of neural networks with abstract interpretation. In *2018 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18, 2018.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL `http://arxiv.org/abs/1412.6572`.

Xiaowei Huang, Marta Kwiatkowska, Sen Wang, and Min Wu. Safety verification of deep neural networks. In *International Conference on Computer Aided Verification*, pp. 3–29. Springer, 2017.

Guy Katz, Clark Barrett, David L Dill, Kyle Julian, and Mykel J Kochenderfer. Reluplex: An efficient smt solver for verifying deep neural networks. In *International Conference on Computer Aided Verification*, pp. 97–117. Springer, 2017.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

Yann LeCun, Corinna Cortes, and Chris Burges. Mnist handwritten digit database, 2010.

Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE Symposium on Security and Privacy (SP)*, pp. 656–672. IEEE, 2019.

Klas Leino, Zifan Wang, and Matt Fredrikson. Globally-robust neural networks. In *International Conference on Machine Learning (ICML)*, 2021.

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, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pp. 506–519, 2017.

Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=Bys4ob-Rb`.

Wenjie Ruan, Min Wu, Youcheng Sun, Xiaowei Huang, Daniel Kroening, and Marta Kwiatkowska. Global robustness evaluation of deep neural networks with provable guarantees for the hamming distance. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pp. 5944–5952. International Joint Conferences on Artificial Intelligence Organization, 7 2019. doi: 10.24963/ijcai.2019/824. URL `https://doi.org/10.24963/ijcai.2019/824`.

Vikash Sehwag, Arjun Nitin Bhagoji, Liwei Song, Chawin Sitawarin, Daniel Cullina, Mung Chiang, and Prateek Mittal. Analyzing the robustness of open-world machine learning. In *Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security*, pp. 105–116, 2019.

Gagandeep Singh, Timon Gehr, Markus Püschel, and Martin Vechev. An abstract domain for certifying neural networks. *Proc. ACM Program. Lang.*, 3(POPL), January 2019.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In Yoshua Bengio and Yann LeCun (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL `http://arxiv.org/abs/1312.6199`.

Vincent Tjeng, Kai Y. Xiao, and Russ Tedrake. Evaluating robustness of neural networks with mixed integer programming. In *International Conference on Learning Representations*, 2019. URL `https://openreview.net/forum?id=HyGIdiRqtm`.

Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. *Advances in Neural Information Processing Systems*, 33, 2020.

Lily Weng, Huan Zhang, Hongge Chen, Zhao Song, Cho-Jui Hsieh, Luca Daniel, Duane Boning, and Inderjit Dhillon. Towards fast computation of certified robustness for relu networks. In *International Conference on Machine Learning*, pp. 5276–5285. PMLR, 2018.

Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning*, pp. 5286–5295. PMLR, 2018.

Eric Wong, Frank R. Schmidt, Jan Hendrik Metzen, and J. Zico Kolter. Scaling provable adversarial defenses. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 84108419, Red Hook, NY, USA, 2018. Curran Associates Inc.

Greg Yang, Tony Duan, J Edward Hu, Hadi Salman, Ilya Razenshteyn, and Jerry Li. Randomized smoothing of all shapes and sizes. In *International Conference on Machine Learning*, pp. 10693–10705. PMLR, 2020a.

Yao-Yuan Yang, Cyrus Rashtchian, Hongyang Zhang, Russ R Salakhutdinov, and Kamalika Chaudhuri. A closer look at accuracy vs. robustness. In *NeurIPS*, 2020b.