The Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)

Robustness Certificates for Sparse Adversarial Attacks by Randomized Ablation

Alexander Levine, Soheil Feizi
University of Maryland, College Park
{alevine0, sfeizi}@cs.umd.edu

Abstract

Recently, techniques have been developed to provably guarantee the robustness of a classifier to adversarial perturbations of bounded L_1 and L_2 magnitudes by using randomized smoothing: the robust classification is a consensus of base classifications on randomly noised samples where the noise is additive. In this paper, we extend this technique to the L_0 threat model. We propose an efficient and certifiably robust defense against sparse adversarial attacks by randomly ablating input features, rather than using additive noise. Experimentally, on MNIST, we can certify the classifications of over 50% of images to be robust to any distortion of at most 8 pixels. This is comparable to the observed empirical robustness of unprotected classifiers on MNIST to modern L_0 attacks, demonstrating the tightness of the proposed robustness certificate. We also evaluate our certificate on ImageNet and CIFAR-10. Our certificates represent an improvement on those provided in a concurrent work (Lee et al. 2019) which uses random noise rather than ablation (median certificates of 8 pixels versus 4 pixels on MNIST; 16 pixels versus 1 pixel on ImageNet.) Additionally, we empirically demonstrate that our classifier is highly robust to modern sparse adversarial attacks on MNIST. Our classifications are robust, in median, to adversarial perturbations of up to 31 pixels, compared to 22 pixels reported as the state-of-the-art defense, at the cost of a slight decrease (around 2.3%) in the classification accuracy. Code and supplementary material is available at https://github.com/alevine0/randomizedAblation/.

Introduction

Adversarial attacks, and defenses against these attacks, have been active topics of research in machine learning in recent years (Szegedy et al. 2013; Carlini and Wagner 2017; Madry et al. 2017). In the case of image classification, given a classifier f, the goal of an adversarial attack on an image $\mathbf{x}$ is to produce an image $\mathbf{x}'$, such that $\mathbf{x}'$ is visually similar to $\mathbf{x}$, but f classifies $\mathbf{x}'$ differently than it classifies $\mathbf{x}$. Assuming that $\mathbf{x}$ is a natural image that was classified correctly, this means that the attacker can produce an image $\mathbf{x}'$ which looks imperceptibly similar to this natural image, but is misclassified by f.

Copyright © 2020, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

When designing or evaluating an adversarial attack, one must choose an objective measure of 'similarity' between two images: more precisely, the goal of the attacker is to minimize $d(\mathbf{x}, \mathbf{x}')$, subject to $f(\mathbf{x}) \neq f(\mathbf{x}')$, where d is a chosen distance metric. Most existing work in adversarial examples has used L_p norms as distance metrics, focusing in particular on L_∞ and L_2 norms (Goodfellow, Shlens, and Szegedy 2015; Szegedy et al. 2013; Madry et al. 2017; Dong et al. 2018; Kurakin, Goodfellow, and Bengio 2018). The L_0 metric, which is simply the number of pixels at which $\mathbf{x}'$ differs from $\mathbf{x}$, has also been the target of adversarial attacks. This metric presents a distinct challenge, because $d(\mathbf{x}, \mathbf{x}')$ is non-differentiable. However, both gradient-based (white-box) attacks (Madry et al. 2017; Papernot et al. 2016a) and zeroth-order (black-box) attacks (Schott et al. 2019) have been proposed under the L_0 attack model. The L_0 attack model is the focus of this paper.

Several practical defenses against adversarial attacks under the L_0 attack model have been proposed in the last couple of years. These methods include defensive distillation (Papernot et al. 2016b), as well as attempts to recover $\mathbf{x}$ from $\mathbf{x}'$ using compressed sensing (Bafna, Murtagh, and Vyas 2018) or generative models (Schott et al. 2019; Meng and Chen 2017). However, as new defenses are proposed, new attacks are also developed for which these defenses are vulnerable (e.g. (Carlini and Wagner 2016)). Experimental demonstrations of a defense's efficacy based on currently existing attacks do not provide a general proof of security. In response, *certifiably robust* classifiers have been developed for adversarial examples for a variety of attack models (Wong and Kolter 2018; Gowal et al. 2018). For these classifiers, given an image $\mathbf{x}$, it is possible to compute a radius ρ such that it is guaranteed that no adversarial example $\mathbf{x}'$ exists within a distance ρ of $\mathbf{x}$. One drawback of many of these certifiable approaches is that they can be computationally expensive since they attempt to minimize $d(\mathbf{x}, \mathbf{x}')$ (or its lower bound) using formal methods.

Recently, a relatively computationally inexpensive family of certifiably robust classifiers have been proposed which employ *randomized smoothing* (Lecuyer et al. 2019; Cohen, Rosenfeld, and Kolter 2019; Li et al. 2018; Salman et al. 2019). This development has mostly been focused on the L_1

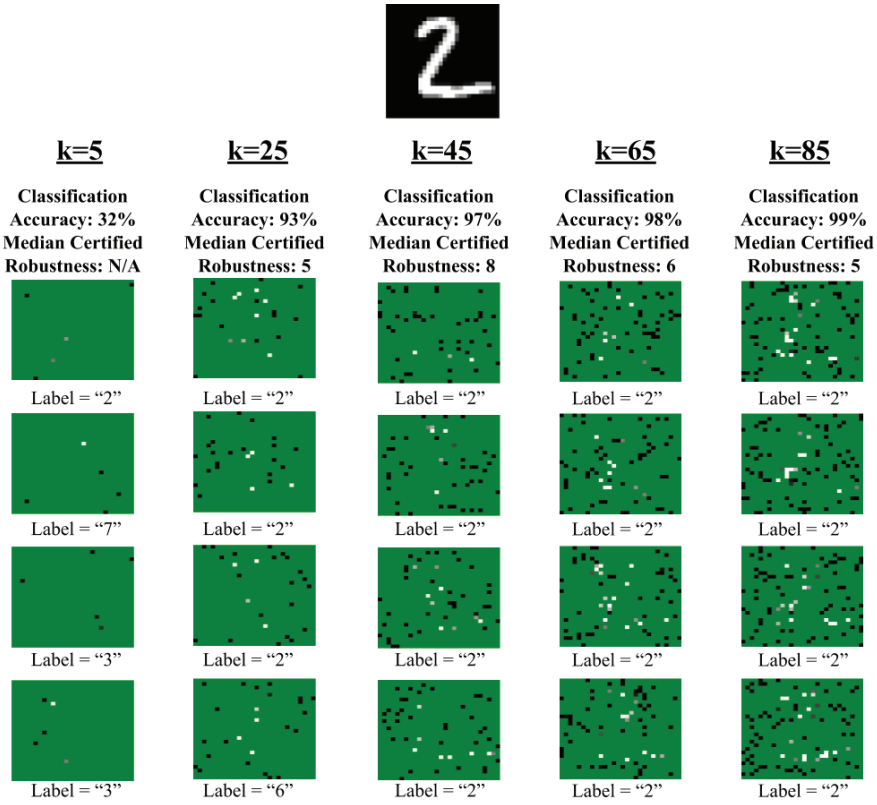

Figure 1: An illustration of our proposed certifiably robust classification scheme on MNIST. At the top, the image to be classified is shown. For randomly ablated images, we retain only k out of 784 total pixels (green pixels in these images are not used in classification). For each value of k, we show four randomly ablated images along with their base classifier labels. For small values of k, the *smoothed* classifier's accuracy in the test set is low ($\sim 32\%$ for $k = 5$) while the accuracy increases for moderate values of k ($\sim 97\%$ for $k = 45$). In each case, we compute the *median certified robustness* for the smoothed classifier of the L_0 attack magnitude that classifications are provably protected against. The median is over the MNIST test set. For example, for $k = 45$, we guarantee the robustness of our proposed method against all L_0 adversarial attacks that perturb 8 or fewer pixels.

and L_2 metrics. Conceptually, these schemes work by repeatedly adding random noise to the image $\mathbf{x}$, in order to create a large set of noised images. A *base classifier* is then used to classify each of these noised samples, and the final robust classification is made by 'majority vote.' The key insight is that, if the magnitude of the noise added to each image is much larger than the distance between $\mathbf{x}$ and a potentially adversarial image $\mathbf{x}'$, then any particular noised image generated from $\mathbf{x}$ could have been generated from $\mathbf{x}'$ with nearly equal likelihood. Then the expected number of 'votes' for each class can only differ between $\mathbf{x}$ and $\mathbf{x}'$ by a bounded amount. Therefore, if we use a statistically sufficient number of random noise samples, and if the observed 'gap' between the number of votes for the top class and the number of 'votes' for any other class at $\mathbf{x}$ is sufficiently large, then we can guarantee with high probability that the robust classification at $\mathbf{x}'$ will be the same as it is at $\mathbf{x}$. Note that the success probability can be made arbitrarily high by adding more noise samples to $\mathbf{x}$ in the smoothing process.

In this work, we develop a certifiably robust classification scheme for the L_0 metric (i.e. sparse adversarial perturbations). To guarantee the robustness of the classification against sparse adversarial attacks, we propose a novel smoothing method based on performing random *ablations* on the input image, rather than adding random noise. In our proposed L_0 smoothing method, for each sample generated from $\mathbf{x}$, a majority of pixels are randomly dropped from the image before the image is given to the base classifier. If a relatively small number ρ of pixels have been adversarially corrupted (which is the case in sparse adversarial attacks), then it is highly likely that none of these pixels are present in a given ablated sample. Then, for the majority of possible random ablations, $\mathbf{x}$ and $\mathbf{x}'$ will give the same ablated image. Therefore, the expected number of votes for each class can only differ between $\mathbf{x}$ and $\mathbf{x}'$ by a bounded amount. Using this, we can prove that with high probability, the smoothed classifier will classify $\mathbf{x}$ robustly against any sparse adversarial attack which is allowed to perturbed certain number of input pixels, provided that the 'gap' between the number of votes for the top class and the number of 'votes' for any other class at $\mathbf{x}$ is sufficiently large. (See Figure 1)

Our ablation method produces significantly larger robustness guarantees compared to a more direct extension of randomized smoothing to the L_0 metric provided in a concur-

rent work by (Lee et al. 2019): see the Discussion section for a comparison of the techniques.

We note that our proposed approach bears some similarities to (Hosseini, Kannan, and Poovendran 2019), in that both works aim to defend against L_0 adversarial attacks by randomly ablating pixels. However, several differences exist: most notably, (Hosseini, Kannan, and Poovendran 2019) presents a *practical* defense with no robustness certificate given. By contrast, the main contribution of this work is a provable guarantee of robustness to adversarial attack.

In summary, our contributions are as follows:

- We develop a novel defense technique against sparse adversarial attacks (threat models that use the L_0 metric) based on randomized ablation.
- We characterize robustness guarantees for our proposed defense against arbitrary sparse adversarial attacks.
- We show the effectiveness of the proposed technique on standard datasets: MNIST, CIFAR-10, and ImageNet.

Preliminaries and Notation

We will use $\mathcal{S}$ to represent the set of possible pixel values in an image. For example, in an 24-bit RGB color image, $\mathcal{S} = \{0, 1, ..., 255\}^3$, while in a binarized black-and-white image, $\mathcal{S} = \{0, 1\}$. We will use $\mathcal{X} = \mathcal{S}^d$ to represent the set of possible images, where d is the number of pixels in each image. Additionally, we will use $\mathcal{S}_{\text{NULL}}$ to represent the set $\mathcal{S} \cup \{\text{NULL}\}$, where NULL is a null symbol representing the absence of information about a pixel, and $\mathcal{X}_{\text{NULL}} = {\mathcal{S}_{\text{NULL}}}^d$ to represent the set of images where some elements in the images may be replaced by the null symbol. Note that NULL is *not* the same as a zero-valued pixel, or black. For example, if $\mathcal{S} = \{0, 1\}$ and $d = 5$, then $[0, 1, 1, 0, 1]^T \in \mathcal{X}$, while $[\text{NULL}, 1, \text{NULL}, 0, 1]^T \in \mathcal{X}_{\text{NULL}}$.

Also, let $[d]$ represent the set of indices $\{1, ..., d\}$, let $\mathcal{H}(d, k) \subseteq \mathcal{P}([d])$ represent all sets of k unique indices in $[d]$, and let $\mathcal{U}(d, k)$ represent the uniform distribution over $\mathcal{H}(d, k)$. (To sample from $\mathcal{U}(d, k)$ is to sample k out of d indices uniformly *without replacement*. For example, an element sampled from $\mathcal{U}(5, 3)$ might be $\{2, 4, 5\}$.)

We define the operation $\textsc{Ablate} \in \mathcal{X} \times \mathcal{H}(d, k) \to \mathcal{X}_{\text{NULL}}$, which takes an image and a set of indices, and outputs the image, with all pixels *except* those in the set replaced with the null symbol NULL. For example, $\textsc{Ablate}([0, 1, 1, 0, 1]^T, \{2, 4, 5\}) = [\text{NULL}, 1, \text{NULL}, 0, 1]^T$

For images $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$, let $\|\mathbf{x} - \mathbf{x}'\|_0$ denote the L_0 distance between $\mathbf{x}$ and $\mathbf{x}'$, defined as the number of pixels at which $\mathbf{x}$ and $\mathbf{x}'$ differ. Note that we are following the convention used by (Carlini and Wagner 2017), where, for a color image, the number of channels in which the images differ at a given pixel location does not matter: any difference at a pixel location (corresponding to an index in $[d]$) counts the same. This differs from (Papernot et al. 2016a), in which channels are counted separately. Also (in a slight abuse of notation) let $\mathbf{x} \ominus \mathbf{x}'$ denote the set of pixel indices at which $\mathbf{x}$ and $\mathbf{x}'$ differ, so that $\|\mathbf{x} - \mathbf{x}'\|_0 = |\mathbf{x} \ominus \mathbf{x}'|$.

Finally, for multiclass classification problems, let c be the number of classes.

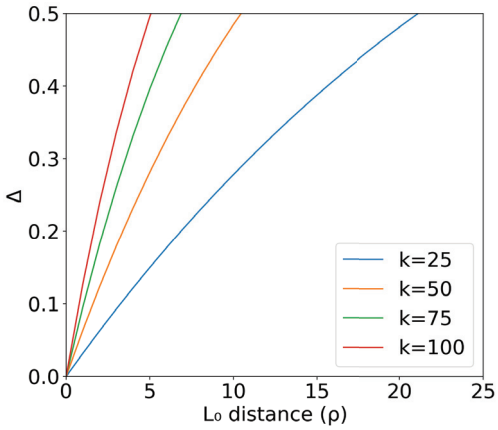

Figure 2: The bounding constant Δ from Theorem 1, shown for MNIST-sized images (d=784). The constant k is the number of pixels retained in each randomly ablated sample.

Certifiably Robust Classification Scheme

First, we note that in this section, we closely follow the notation of (Cohen, Rosenfeld, and Kolter 2019), using appropriate analogs between the L_2 smoothing scheme of that work, and the proposed L_0 ablation scheme of this work. In particular, let $f \in \mathcal{X}_{\text{NULL}} \to [c]$ denote a *base classifier*, which is trained to classify images with some pixels ablated. Let $g \in \mathcal{X} \to [c]$ represent a *smoothed classifier*, defined as:

$$g(\mathbf{x}) = \arg\max_i \left[\Pr_{\mathcal{T} \sim \mathcal{U}(d,k)} (f(\textsc{Ablate}(\mathbf{x}, \mathcal{T})) = i) \right] \quad (1)$$

where k is the *retention constant*; i.e., the number of pixels retained (not ablated) from $\mathbf{x}$. In other words, $g(\mathbf{x})$ denotes the class *most likely to be returned* if we first randomly ablate all but k pixels from $\mathbf{x}$ and then classify the resulting image with the base classifier f. To simplify notation, we will let $p_i(\mathbf{x})$ denote the probability that, after ablation, f returns the class i:

$$p_i(\mathbf{x}) = \Pr_{\mathcal{T} \sim \mathcal{U}(d,k)} (f(\textsc{Ablate}(\mathbf{x}, \mathcal{T})) = i) \,. \quad (2)$$

Thus, $g(\mathbf{x})$ can be defined simply as $\arg\max_i [p_i(\mathbf{x})]$.

We first prove the following general theorem, which can be used to develop a variety of related robustness certificates.

Theorem 1. *For images $\mathbf{x}, \mathbf{x}'$, with $\|\mathbf{x} - \mathbf{x}'\|_0 \leq \rho$, for all classes $i \in [c]$:*

$$|p_i(\mathbf{x}') - p_i(\mathbf{x})| \leq \Delta \quad (3)$$

where

$$\Delta = 1 - \frac{\binom{d-\rho}{k}}{\binom{d}{k}}. \quad (4)$$

See Figure 2 for a plot of how the constant Δ scales with k and ρ. We present a short proof of Theorem 1 here:

Proof. Recall that (with $\mathcal{T} \sim \mathcal{U}(d, k)$):

$$\begin{aligned} p_i(\mathbf{x}) &= \Pr(f(\textsc{Ablate}(\mathbf{x}, \mathcal{T})) = i) \\ p_i(\mathbf{x}') &= \Pr(f(\textsc{Ablate}(\mathbf{x}', \mathcal{T})) = i) \end{aligned} \quad (5)$$

By the law of total probability:

$$\begin{aligned} &p_i(\mathbf{x}) = \\ &\Pr([f(\textsc{Ablate}(\mathbf{x}, \mathcal{T})) = i] \wedge [\mathcal{T} \cap (\mathbf{x} \ominus \mathbf{x}') = \emptyset]) + \\ &\Pr([f(\textsc{Ablate}(\mathbf{x}, \mathcal{T})) = i] \wedge [\mathcal{T} \cap (\mathbf{x} \ominus \mathbf{x}') \neq \emptyset]) \\ &p_i(\mathbf{x}') = \\ &\Pr([f(\textsc{Ablate}(\mathbf{x}', \mathcal{T})) = i] \wedge [\mathcal{T} \cap (\mathbf{x} \ominus \mathbf{x}') = \emptyset]) + \\ &\Pr([f(\textsc{Ablate}(\mathbf{x}', \mathcal{T})) = i] \wedge [\mathcal{T} \cap (\mathbf{x} \ominus \mathbf{x}') \neq \emptyset]) \end{aligned} \tag{6}$$

Note that if $\mathcal{T} \cap (\mathbf{x} \ominus \mathbf{x}') = \emptyset$, then $\mathbf{x}$ and $\mathbf{x}'$ are identical at all indices in $\mathcal{T}$. Then in this case, $\textsc{Ablate}(\mathbf{x}, \mathcal{T})) = \textsc{Ablate}(\mathbf{x}', \mathcal{T}))$, which implies:

$$\begin{aligned} &\Pr(f(\textsc{Ablate}(\mathbf{x}, \mathcal{T})) = i \mid \mathcal{T} \cap (\mathbf{x} \ominus \mathbf{x}') = \emptyset) = \\ &\Pr(f(\textsc{Ablate}(\mathbf{x}', \mathcal{T})) = i \mid \mathcal{T} \cap (\mathbf{x} \ominus \mathbf{x}') = \emptyset) \end{aligned} \tag{7}$$

Multiplying both sides of (7) by $\Pr(\mathcal{T} \cap (\mathbf{x} \ominus \mathbf{x}') = \emptyset)$ gives:

$$\begin{aligned} &\Pr([f(\textsc{Ablate}(\mathbf{x}, \mathcal{T})) = i] \wedge [\mathcal{T} \cap (\mathbf{x} \ominus \mathbf{x}') = \emptyset]) = \\ &\Pr([f(\textsc{Ablate}(\mathbf{x}', \mathcal{T})) = i] \wedge [\mathcal{T} \cap (\mathbf{x} \ominus \mathbf{x}') = \emptyset]) \end{aligned} \tag{8}$$

Substituting (8) into (6) and rearranging yields:

$$\begin{aligned} &p_i(\mathbf{x}') = p_i(\mathbf{x}) - \\ &\Pr([f(\textsc{Ablate}(\mathbf{x}, \mathcal{T})) = i] \wedge [\mathcal{T} \cap (\mathbf{x} \ominus \mathbf{x}') \neq \emptyset]) + \\ &\Pr([f(\textsc{Ablate}(\mathbf{x}', \mathcal{T})) = i] \wedge [\mathcal{T} \cap (\mathbf{x} \ominus \mathbf{x}') \neq \emptyset]) \end{aligned} \tag{9}$$

Because probabilities are non-negative, this gives:

$$\begin{aligned} &p_i(\mathbf{x}) - \\ &\Pr([f(\textsc{Ablate}(\mathbf{x}, \mathcal{T})) = i] \wedge [\mathcal{T} \cap (\mathbf{x} \ominus \mathbf{x}') \neq \emptyset]) \\ &\leq p_i(\mathbf{x}') \leq \\ &p_i(\mathbf{x}) + \\ &\Pr([f(\textsc{Ablate}(\mathbf{x}', \mathcal{T})) = i] \wedge [\mathcal{T} \cap (\mathbf{x} \ominus \mathbf{x}') \neq \emptyset]) \end{aligned} \tag{10}$$

By the conjunction rule, this implies:

$$\begin{aligned} &p_i(\mathbf{x}) - \Pr(\mathcal{T} \cap (\mathbf{x} \ominus \mathbf{x}') \neq \emptyset) \\ \leq &p_i(\mathbf{x}') \leq \\ &p_i(\mathbf{x}) + \Pr(\mathcal{T} \cap (\mathbf{x} \ominus \mathbf{x}') \neq \emptyset) \end{aligned} \tag{11}$$

Note that:

$$\begin{aligned} &\Pr(\mathcal{T} \cap (\mathbf{x} \ominus \mathbf{x}') \neq \emptyset) = \\ &1 - \Pr(\mathcal{T} \cap (\mathbf{x} \ominus \mathbf{x}') = \emptyset) = 1 - \frac{\binom{d - |\mathbf{x} \ominus \mathbf{x}'|}{k}}{\binom{d}{k}} \end{aligned} \tag{12}$$

Where the last equality follows because $\mathcal{T}$ is an uniform choice of k elements from d: there are $\binom{d}{k}$ total ways to make this selection, $\binom{d - |\mathbf{x} \ominus \mathbf{x}'|}{k}$ of which contain no elements from $(\mathbf{x} \ominus \mathbf{x}')$. Then:

$$\begin{aligned} \Pr(\mathcal{T} \cap (\mathbf{x} \ominus \mathbf{x}') \neq \emptyset) &= 1 - \frac{\binom{d - |\mathbf{x} \ominus \mathbf{x}'|}{k}}{\binom{d}{k}} \\ &= 1 - \frac{\binom{d - \|\mathbf{x} - \mathbf{x}'\|_0}{k}}{\binom{d}{k}} \leq 1 - \frac{\binom{d-\rho}{k}}{\binom{d}{k}} = \Delta \end{aligned} \tag{13}$$

Combining inequalities (13) and (11) gives the statement of Theorem 1. □

Practical Robustness Certificates

Depending on the architecture of the base classifier, it may be infeasable to directly compute $p_i(\mathbf{x})$, and therefore to compute $g(\mathbf{x})$. However, we can instead generate a representative sample from $\mathcal{U}(d, k)$, in order to bound $p_i(\mathbf{x})$ with high confidence. In particular, let $\underline{p_i(\mathbf{x})}$ represent a lower bound on $p_i(\mathbf{x})$, with $(1 - \alpha)$ confidence, and let $\overline{p_i(\mathbf{x})}$ represent a similar upper bound. We first develop a certificate analogous for the L_0 attack to the certificate presented in (Cohen, Rosenfeld, and Kolter 2019):

Corollary 1. *For images $\mathbf{x}, \mathbf{x}'$, with $\|\mathbf{x} - \mathbf{x}'\|_0 \leq \rho$, if:*

$$\underline{p_i(\mathbf{x})} - \Delta > 0.5 \tag{14}$$

then, with probability at least $1 - \alpha$:

$$g(\mathbf{x}') = i \tag{15}$$

Proof. With probability at least $1 - \alpha$:

$$.5 < \underline{p_i(\mathbf{x})} - \Delta \leq p_i(\mathbf{x}) - \Delta \leq p_i(\mathbf{x}') \tag{16}$$

where the final inequality is from Theorem 1. Then $g(\mathbf{x}') = i$ from the definition of g. □

This bound applies directly to the true population value of $g(\mathbf{x}')$, not necessarily to an empirical estimate of $g(\mathbf{x}')$. Following (Cohen, Rosenfeld, and Kolter 2019), we therefore use a separate sampling procedure to estimate the value of the classifier $g(.)$, which itself has a bounded failure rate independent from the failure rate of the certificate, and which may abstain from classification if the top class probabilities are too similar to distinguish based on the samples. Note that by using a large number of samples, this estimation error can be made arbitrarily small. In fact, because Corollary 1 is directly analogous to the condition for L_2 robustness presented in (Cohen, Rosenfeld, and Kolter 2019), we borrow both the empirical classification and the empirical certification procedures from that paper wholesale. We refer the reader to that work for details: it is sufficient to say that with these procedures, we can bound $\underline{p_i(\mathbf{x})}$ with $(1 - \alpha)$ confidence and also estimate $g(\mathbf{x}')$ with $(1 - \alpha)$ confidence. This is the procedure we use in our experiments.

Alternatively, one can instead use a certificate analogous to the certificate presented in (Lecuyer et al. 2019).

Corollary 2. *For images $\mathbf{x}, \mathbf{x}'$, with $\|\mathbf{x} - \mathbf{x}'\|_0 \leq \rho$, if:*

$$\underline{p_i(\mathbf{x})} - \Delta > \arg\max_{k \neq i} \overline{p_k(\mathbf{x})} + \Delta \tag{17}$$

then, with probability at least $1 - \alpha$:

$$g(\mathbf{x}') = i. \tag{18}$$

Proof. For each $k \neq i$:

$$\begin{aligned} p_k(\mathbf{x}') &\leq p_k(\mathbf{x}) + \Delta \leq \overline{p_k(\mathbf{x})} + \Delta \leq \arg\max_{k \neq i} \overline{p_k(\mathbf{x})} + \Delta \\ &< \underline{p_i(\mathbf{x})} - \Delta \leq p_i(\mathbf{x}) - \Delta \leq p_i(\mathbf{x}') \end{aligned} \tag{19}$$

where the first and last inequalities are from Theorem 1. □

In a multi-class setting, Corollary 2 might appear to give a tighter certificate bound. However, the upper and lower bounds on $p_j(\mathbf{x})$ must hold simultaneously for all j with a total failure rate of $(1 - \alpha)$. This can lead to greater estimation error if the number of classes c is large.

Architectural and training considerations

Similar to existing works on smoothing-based certified adversarial robustness, we train our base classifier f on noisy images (i.e. ablated images), rather than training g directly. For performance reasons, during training, we ablate the same pixels from all images in a minibatch. We use the same retention constant k during training as at test time.

Encoding $\mathcal{S}_{\text{NULL}}$. We use standard CNN-based architectures for the classifier $f(.)$. However, this presents an architectural challenge: we need to be able to represent the absence of information at a pixel (the symbol NULL), as distinct from any color that can normally be encoded. Additionally, we would like the encoding of NULL to be equally far from every possible encodable color, so that the network is not biased towards treating it as one color moreso than another. To achieve these goals, we encode images as follows: for greyscale images where pixels in $\mathcal{S}$ are floating point values between zero and one (i.e. $\mathcal{S} = [0, 1]$), we encode $s \in \mathcal{S}$ as the tuple $(s, 1 - s)$, and then encode NULL as $(0, 0)$. Practically, this means that we double the number of color channels from one to two, with one channel representing the original image and the other channel representing its inverse. Then, NULL is represented as zero on both channels: this is distinct from grey $(0.5, 0.5)$, white $(1, 0)$, or black $(0, 1)$. Notably, the values over the channels add up to one for a pixel representing any color, while it adds up to zero for a null pixel. For color images, we use the same encoding technique increasing the number of channels from 3 to 6. The resulting channels are then $(\text{red}, \text{green}, \text{blue}, 1 - \text{red}, 1 - \text{green}, 1 - \text{blue})$, while NULL is encoded as $(0, 0, 0, 0, 0, 0)$.[1]

Results

In this section, we provide experimental results of the proposed method on MNIST, CIFAR-10, and ImageNet. When reporting results, we refer to the following quantities:

- The *certified robustness* of a particular image $\mathbf{x}$ is the maximum ρ for which we can certify (with probability at least $1 - \alpha$) that the smoothed classifier $g(\mathbf{x}')$ will return the *correct* label where $\mathbf{x}'$ is any adversarial perturbation of $\mathbf{x}$ such that $\|\mathbf{x} - \mathbf{x}'\|_0 \leq \rho$. If the unperturbed classification $g(\mathbf{x})$ is itself incorrect, we define the certified robustness as N/A (Not Applicable).
- The *certified accuracy at* ρ on a dataset is the fraction of images in the dataset with *certified robustness* of at least ρ. In other words, it is the guaranteed accuracy of the classifier $g(.)$, if all images are corrupted with any L_0 adversarial attack of measure up to ρ.
- The *median certified robustness* on a dataset is the median value of the *certified robustness* across the dataset. Equivalently, it is the maximum ρ for which the *certified accuracy at* ρ is at least 0.5. When computing this median, images which $g(.)$ misclassifies when unperturbed (i.e., *certified robustness* is N/A) are counted as having $-\infty$ certified robustness. For example, if the robustness certificates of images in a dataset are $\{\text{N/A}, \text{N/A}, 1, 2, 3\}$, the *median certified robustness* is 1, not 2.
- The *classification accuracy* on a dataset is the fraction of images on which our empirical estimation of $g(.)$ returns the correct class label, and does not abstain.
- The *empirical adversarial attack magnitude* of a particular image $\mathbf{x}$ is the minimum ρ for which an adversarial attack can find an adversarial example $\mathbf{x}'$ such that $\|\mathbf{x} - \mathbf{x}'\|_0 \leq \rho$, and such that our empirical classification procedure misclassifies or abstains on $\mathbf{x}'$.
- The *median adversarial attack magnitude* on a dataset is the median value of the *empirical adversarial attack magnitude* across the dataset.

Unless otherwise stated, the uncertainty α is 0.05, and 10,000 randomly-ablated samples are used to make each prediction. The empirical estimation procedure we use to generate certificates, from (Cohen, Rosenfeld, and Kolter 2019), requires two sampling steps: the first to identify the majority class i, and the second to bound $\underline{p_i(\mathbf{x})}$. We use 1,000 and 10,000 samples, respectively, for these two steps.

Results on MNIST

We first tested our robust classification scheme on MNIST, using a simple CNN model as the base classifier (see supplementary material for architectural details.) Results are presented in Table 1. We varied the number of retained pixels k in each sample: note that for small k, certified robustness and accuracy both increase as k increases. However, after a certain threshold, here achieved at $k = 45$, certified robustness starts to decrease with k, while classification accuracy continues to increase. This can be understood by considering Figure 2: For larger k, the bounding constant Δ grows considerably faster with the L_0 distance ρ. In other words, a larger fraction of ablated samples must be classified correctly to achieve the same certified robustness. For small k, the fraction of ablated samples classified correctly increases sufficiently quickly with k to counteract this effect; however, after a certain point, it is no longer beneficial to increase k because a large majority of samples are already classified correctly by the base classifier (For example, see Figure 1).

We also tested the empirical robustness of our classifier to an L_0 adversarial attack. Specifically, we chose to use the black-box *Pointwise attack* proposed by (Schott et al. 2019). We choose a black-box attack because comparisons to other robust classifiers using gradient-based attacks (such as the L_0 attack proposed by (Carlini and Wagner 2017)) may be somewhat asymmetric since our smoothed classifier is non-differentiable (because the base classifier's output is discretized.) While (Salman et al. 2019) does propose a gradient-based scheme for attacking L_2-smoothed classifiers which are similarly non-differentiable, adapting such a scheme would be a non-trivial departure from the existing L_0 Carlini-Wagner attack, precluding a direct comparison to other robust classifiers. By contrast, a practical reason

[1] On CIFAR-10, we scaled colors between 0 and 1 when using this encoding. On ImageNet, we normalized each channel to have mean 0 and standard deviation 1 before applying this encoding: in this case, the NULL symbol is still distinct, although it is not equidistant from all other colors.

Retained pixels k	Classification accuracy (Percent abstained)	Median certified robustness
5	32.32% (5.65%)	N/A
10	74.90% (5.08%)	0
15	86.09% (2.82%)	0
20	90.29% (1.81%)	3
25	93.05% (1.02%)	5
30	94.68% (0.77%)	7
35	95.40% (0.66%)	7
40	96.27% (0.52%)	8
45	**96.72% (0.45%)**	**8**
50	97.16% (0.32%)	7
55	97.41% (0.34%)	7
60	97.78% (0.18%)	7
65	98.05% (0.15%)	6
70	98.18% (0.20%)	6
75	98.28% (0.20%)	6
80	98.37% (0.12%)	5
85	98.57% (0.12%)	5
90	98.58% (0.16%)	5
95	98.73% (0.11%)	5
100	98.75% (0.16%)	4

Table 1: Robustness certificates on MNIST, using different numbers of retained pixels (k). The maximum median certified robustness on the MNIST test set is achieved when using $k = 40$ or $k = 45$ retained pixels: because $k = 45$ gives better classification accuracy, we use this model (highlighted in bold) when evaluating against adversarial attacks.

we choose the Pointwise Attack is that the reference implementation of the attack is available as part of the Foolbox package (Rauber, Brendel, and Bethge 2017), meaning that we can directly compare our results to that of (Schott et al. 2019), without any concerns about implementation details. We note that (Schott et al. 2019) reports a median adversarial attack magnitude of 9 pixels for an unprotected CNN model on MNIST, which is comparable to the *mean* adversarial attack magnitude of 8.5 reported for the L_0 Carlini-Wagner attack. This suggests that the attack is comparably effective. Results are presented in Table 2. Note that our model appears to be significantly more robust to L_0 attack than any of the models tested by (Schott et al. 2019), at a slight cost of classification accuracy (We would anticipate this tradeoff, see (Tsipras et al. 2019).) Also note that while there is a gap between the median certified lower bound for the magnitude of any attack, 8 pixels, and the empirical upper bound given by an extant attack, 31 pixels, these quantities are at least in the same order of magnitude, indicating that our certificate is a non-trivial guarantee. See Figure 3 for examples of adversarial attacks on our classifier.

Results on CIFAR-10

We implemented our technique on CIFAR-10 using ResNet18 (with the number of input channels increased to 6) as a base classifier; see Table 3 for our robustness certificates as a function of k. The median certified robustness is somewhat smaller than for MNIST: however, this

Model	Class. acc.	Median adv. attack mag.
CNN	99.1%	9.0
Binarized CNN	98.5%	11.0
Nearest Neighbor	96.9%	10.0
L_∞-Robust (Madry et al. 2017)	98.8%	4.0
(Schott et al. 2019)	99.0%	16.5
Binarized (Schott et al. 2019)	99.0%	22.0
Our model ($k = 45$)	**96.7%**	**31.0**

Table 2: Median adversarial attack magnitude on MNIST using the Pointwise attack from (Schott et al. 2019), taking the best attack on each image from 10 random restarts. Note that all values except for our model are taken directly from (Schott et al. 2019). For every evaluation performed by the black-box attack, 10,000 ablated samples were used to calculate class scores of our model: this was to ensure stability of the evaluated scores. Additionally, causing our model to abstain from classifying was counted as a successful attack, even if the correct class score was still marginally highest. Because the black-box attack performs a large number of classifications, and each of these classifications required 10,000 evaluations of the base classifier, we used only a subset of the MNIST test set, consisting of 275 images.

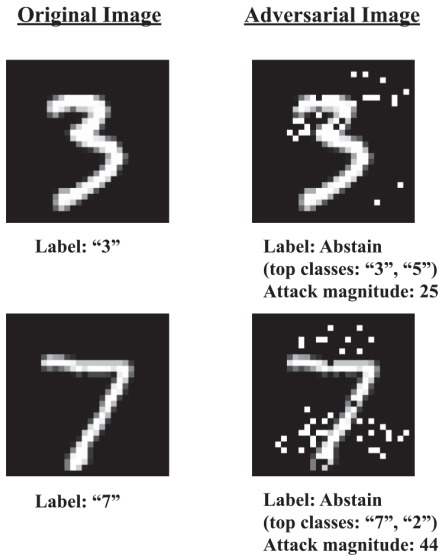

Figure 3: Adversarial examples to our classifier on MNIST. Note that because we consider the classifier abstaining to be a successful attack, these adversarial examples are in fact on the boundary between classes, rather than being entirely misclassified.

is in line with the performance of empirical attacks. For example, the L_0 attack proposed by (Carlini and Wagner 2017) achieves a mean adversarial attack magnitude of 8.5 pixels on MNIST and 5.9 pixels on CIFAR-10. This suggests that CIFAR-10 samples are more vulnerable to L_0 adversarial attacks compared to the MNIST ones. Intuitively,

Retained pixels k	Classification accuracy (Percent abstained)	Median certified robustness
25	68.41% (1.76%)	6
50	74.21% (1.19%)	7
75	**78.25% (0.93%)**	**7**
100	80.91% (0.86%)	6
125	83.25% (0.60%)	5
150	85.22% (0.53%)	4

Table 3: Robustness certificates on CIFAR-10, using different numbers of retained pixels (k), and using ResNet18 (He et al. 2016) as the base classifier. Note that without smoothing, the base implementation of an unprotected ResNet18 classifier which we used (Liu 2019) has a classification accuracy of 93.02% on CIFAR-10.

Retained pixels k	Base classifier training accuracy	Base classifier test accuracy
25	83.16%	57.72%
50	96.63%	68.29%
75	99.33%	74.08%
100	99.76%	77.88%
125	99.91%	80.48%
150	99.95%	83.16%

Table 4: Accuracy of the base classifier f in CIFAR-10 experiments, on training versus test data, using ResNet18. Note that the base classifier significantly overfits to the training data. (Training accuracies are averaged over the final epoch of training.)

this is because CIFAR-10 images are both visually complex and low-resolution, so that each pixel carries a large amount of information regarding the classification label. Also note that the classification accuracy on unperturbed images is somewhat reduced. For example, in a model using $k = 150$, the median certified robustness is 4 pixels, and the classifier accuracy is 85.22%. The trade-off between accuracy and robustness is also more pronounced. However, it is not unusual for practical L_0 defenses to achieve accuracy below 90% on CIFAR-10 (Meng and Chen 2017; Xu, Evans, and Qi 2017): our defense may therefore still prove to be usable.

One phenomenon which we encountered when applying our technique to CIFAR-10 was over-fitting of the base classifier (see Table 4), which was unexpected because during the training, the classifier is always exposed to new random ablations of the training data. However, the network was still able to memorize the training data, despite never being exposed to the complete images. While interpolation of even randomly labeled training data is a known phenomenon in deep learning (Zhang et al. 2017), we were surprised to see that over-fitting may happen on ablated images, where a particular ablation is likely never repeated in training. In order to better understand this, we use a model trained on a higher-capacity network architecture, ResNet50. The results for the base classifier are given in Table 5. Surprisingly, increasing network capacity decreased the generalization gap slightly

Retained pixels k	Base classifier training accuracy	Base classifier test accuracy
25	83.89%	57.58%
50	96.91%	69.45%
75	99.09%	75.22%
100	99.66%	79.54%
125	99.78%	81.83%
150	99.92%	84.43%

Table 5: Accuracy of the base classifier f in CIFAR-10 experiments, on training versus test data, using ResNet50. Note that the base classifier significantly overfits to the training data: however, for $k > 25$, this higher-capacity model overfits less than ResNet18.

for $k \geq 50$ (Note that because the improvement to the base classifier is only marginal, and because ResNet50 is substantially more computationally intensive to use as a base classifier to classify 10,000 ablated samples per image, we opted to compute certificates using the ResNet18 model).

Results on ImageNet

We implemented our technique on ImageNet using ResNet50 (again with the number of input channels increased to 6) as a base classifier; see Table 6 for our robustness certificates as a function of k. For testing, we used a random subset of 400 images from the ILSVRC2012 validation set. Note that ImageNet classification is a 1,000-class problem: here we consider only top-1 accuracy. Because these top-1 accuracies are only moderately above 50 percent, the calculation of the median certified robustness is skewed by relatively large fraction of misclassified points: on the points which are correctly classified, the certificates can be considerably larger. For example, at $k = 1000$, if we consider only the 61% of images which are certified for the correct class, the median certificate is *33 pixels*. Similarly, considering only images with certificates other than 'N/A', the median certificates for $k = 500$ and $k = 2000$ are 63 pixels and 16 pixels, respectively.

Retained pixels k	Classification accuracy (Percent abstained)	Median certified robustness
500	52.75% (1.75%)	0
1000	**61.00% (0.00%)**	**16**
2000	62.50% (1.75%)	11

Table 6: Robustness certificates on ImageNet, using different numbers of retained pixels k, and using ResNet50 (He et al. 2016) as the base classifier. For ImageNet, $d = 224 \times 224$. Note that without smoothing, the base implementation of an unprotected ResNet50 classifier can be trained on ImageNet to a top-1 accuracy of 76.15% (Paszke et al. 2017).

Discussion

Comparison to (Lee et al. 2019)

In a concurrent work, (Lee et al. 2019) also present a randomized-smoothing based robustness certification

scheme for the L_0 metric. In this scheme, each pixel is retained with a fixed probability κ and is otherwise assigned to a *random* value from the remaining possible pixel values in $\mathcal{S}$. Note that there is no NULL in this scheme. As a consequence, the base classifier lacks explicit information about *which* pixels are retained from the original image, and which have been randomized. The resulting scheme has considerably lower median certified robustness on the datasets tested in both works[2] (Table 7):

Dataset	Median certified robustness (pixels) (Lee et al. 2019)	Median certified robustness (pixels) (our model)
MNIST	4	**8**
ImageNet	1	**16**

Table 7: Comparison of robustness certificates in (Lee et al. 2019) and in this work, using the optimal choices of hyperparameters tested in each work. Numbers for (Lee et al. 2019) are derived from those reported in that work. Note that for ImageNet, (Lee et al. 2019) considers each color channel as a separate pixel: therefore the median image is robust to distortion in only *one channel* of one pixel. By contrast, our model is robust to distortions in *all channels* in 16 pixels (or, in the limiting case, one channel in 16 pixels).

To illustrate quantitatively how our robust classifier obtains more information from each ablated sample than is available in the *randomly noised* samples in (Lee et al. 2019), let us consider images of ImageNet scale. Because (Lee et al. 2019) considers each color channel as a separate pixel when computing certificates, we will use $\mathcal{S} = \{0, ..., 255\}$, and $d = 3 * 224 * 224$. Using (Lee et al. 2019)'s certificate scheme, in order to certify for one pixel of robustness with $\kappa = 0.1$ probability of pixel retention, we would need to accurately classify noised images with probability $p_i(\mathbf{x}) = .596$. Meanwhile, using our ablation scheme, in order to certify one pixel of robustness by correctly classifying same fraction ($p_i(\mathbf{x}) = .596$) of ablated images, we can retain at most $k = 14521$ pixels. This is 9.6% of pixels, slightly fewer than the expected number retained in (Lee et al. 2019)'s scheme.

However, we will now calculate the *mutual information* between each ablated/noised image and the original image for each scheme: this is the expected number of bits of information about the original image which are obtained from observing the ablated/noised image. For illustrative purposes, we will make the simplifying assumption that the dataset overall is uniformly distributed (while this is obviously not true for image classification, it is a reasonable assumption in other classification tasks.) In our scheme, we have simply

$$I_{\text{ablate}} = \log_2 |\mathcal{S}| * k = 8 * k = 116168 \text{ bits.} \quad (20)$$

[2](Lee et al. 2019) uses a similar scheme to ours to derive an empirical bound on $p_i(\mathbf{x})$; however, that work uses 100 samples to select i and 100,000 samples to bound it, and reports bounds with 99.9% confidence ($\alpha = .001$). In order to provide a fair comparison, we repeated our certifications on MNIST and ImageNet (for optimized values of k) using these empirical certification parameters. This did not change the median robustness certificates.

Each of the k retained pixels provides 8 bits of information. However, in the noising scheme from (Lee et al. 2019), we instead have:

$$\begin{aligned} &I_{\text{Lee et al.}} \\ &= d\left(\log_2 |\mathcal{S}| + \kappa \log_2 \kappa + (1-\kappa) \log_2 \frac{1-\kappa}{|\mathcal{S}|-1}\right) \quad (21) \\ &\approx 50590.4 \text{ bits.} \end{aligned}$$

Therefore, despite using slightly fewer pixels from the original image, over twice the amount of information about the original image is available in our scheme when making each ablated classification. (A derivation of Equation 21 is provided in the supplementary material.)

Alternative encodings of $\mathcal{S}_{\text{NULL}}$

The multichannel encoding of $\mathcal{S}_{\text{NULL}}$ described above, while theoretically well-motivated, is not the only possible encoding scheme. In fact, for MNIST and CIFAR-10, we tested a somewhat simpler encoding for the NULL symbol: we simply used the mean pixel value on the training set, similarly to the practical defense proposed by (Hosseini, Kannan, and Poovendran 2019). We tested using the optimal values of k from the Results section above ($k = 45$ for MNIST and $k = 75$ for CIFAR-10). This resulted in only marginally decreased accuracy and certificate sizes (Table 8):

$\mathcal{S}_{\text{NULL}}$ encoding	Classification acc. (Pct. abstained)	Median certified robustness
MNIST		
Multichannel	**96.72% (0.45%)**	**8**
Mean	96.27% (0.43%)	7
CIFAR-10		
Multichannel	**78.25% (0.93%)**	**7**
Mean	77.71% (1.05%)	**7**

Table 8: Accuracy and robustness using different encoding schemes for $\mathcal{S}_{\text{NULL}}$.

To understand this, note that the *mean* pixel value (grey in both datasets) is not necessarily a *common* value: it is still possible to distinguish which pixels are ablated (Figure 4).

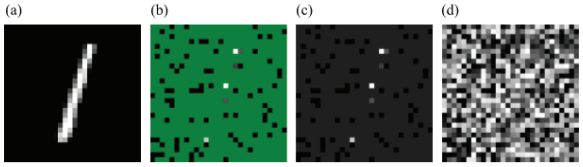

Figure 4: (a) An image from MNIST. (b) The image with $k = 85$ pixels ablated, with a unique NULL encoding. (c) The same image with NULL encoded as the mean pixel value (dark grey). Note that both black and white pixels are still distinguishable. (d) If we replace ablated pixels with random noise, the image is no longer easily distinguishable.

Conclusion

In this paper, we introduced a novel smoothing-based certifiably robust classification method against sparse adversarial attacks, in which the adversary can perturb a certain number features in input samples. Our method, which is modeled after randomised smoothing methods for certifiably robust classification for L_1 and L_2 attack models, was shown to produce non-trivial robustness certificates on MNIST, CIFAR-10, and ImageNet, and to be an effective empirical defense against L_0 attacks on MNIST.

Acknowledgements

This work was supported in part by NSF award CDS&E:1854532 and award HR00111990077.

References

Bafna, M.; Murtagh, J.; and Vyas, N. 2018. Thwarting adversarial examples: An l_0-robust sparse fourier transform. In *Advances in Neural Information Processing Systems*, 10075–10085.

Carlini, N., and Wagner, D. 2016. Defensive distillation is not robust to adversarial examples. *arXiv preprint arXiv:1607.04311*.

Carlini, N., and Wagner, D. 2017. Towards evaluating the robustness of neural networks. In *2017 38th IEEE Symposium on Security and Privacy (SP)*, 39–57. IEEE.

Cohen, J.; Rosenfeld, E.; and Kolter, Z. 2019. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, 1310–1320.

Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; and Li, J. 2018. Boosting adversarial attacks with momentum. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9185–9193. IEEE.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and harnessing adversarial examples. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015*.

Gowal, S.; Dvijotham, K.; Stanforth, R.; Bunel, R.; Qin, C.; Uesato, J.; Mann, T.; and Kohli, P. 2018. On the effectiveness of interval bound propagation for training verifiably robust models. *arXiv preprint arXiv:1810.12715*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hosseini, H.; Kannan, S.; and Poovendran, R. 2019. Dropping pixels for adversarial robustness. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 0–0.

Kurakin, A.; Goodfellow, I. J.; and Bengio, S. 2018. Adversarial examples in the physical world. In *Artificial Intelligence Safety and Security*. Chapman and Hall/CRC. 99–112.

Lecuyer, M.; Atlidakis, V.; Geambasu, R.; Hsu, D.; and Jana, S. 2019. Certified robustness to adversarial examples with differential privacy. In *2019 2019 IEEE Symposium on Security and Privacy (SP)*, 726–742. Los Alamitos, CA, USA: IEEE Computer Society.

Lee, G.-H.; Yuan, Y.; Chang, S.; and Jaakkola, T. S. 2019. Tight certificates of adversarial robustness for randomly smoothed classifiers. *arXiv preprint arXiv:1906.04948*.

Li, B.; Chen, C.; Wang, W.; and Carin, L. 2018. Second-order adversarial attack and certifiable robustness. *arXiv preprint arXiv:1809.03113*.

Liu, K. 2019. 95.16% on cifar10 with pytorch. https://github.com/kuangliu/pytorch-cifar.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*.

Meng, D., and Chen, H. 2017. Magnet: a two-pronged defense against adversarial examples. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 135–147. ACM.

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

Papernot, N.; McDaniel, P.; Wu, X.; Jha, S.; and Swami, A. 2016b. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE Symposium on Security and Privacy (SP)*, 582–597. IEEE.

Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lin, Z.; Desmaison, A.; Antiga, L.; and Lerer, A. 2017. Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*.

Rauber, J.; Brendel, W.; and Bethge, M. 2017. Foolbox: A python toolbox to benchmark the robustness of machine learning models. *arXiv preprint arXiv:1707.04131*.

Salman, H.; Yang, G.; Li, J.; Zhang, P.; Zhang, H.; Razenshteyn, I.; and Bubeck, S. 2019. Provably robust deep learning via adversarially trained smoothed classifiers. *arXiv preprint arXiv:1906.04584*.

Schott, L.; Rauber, J.; Bethge, M.; and Brendel, W. 2019. Towards the first adversarially robust neural network model on mnist. In *Seventh International Conference on Learning Representations (ICLR 2019)*, 1–16.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

Tsipras, D.; Santurkar, S.; Engstrom, L.; Turner, A.; and Madry, A. 2019. Robustness may be at odds with accuracy. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

Wong, E., and Kolter, Z. 2018. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning*, 5283–5292.

Xu, W.; Evans, D.; and Qi, Y. 2017. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155*.

Zhang, C.; Bengio, S.; Hardt, M.; Recht, B.; and Vinyals, O. 2017. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.