Confidence-aware Training of Smoothed Classifiers for Certified Robustness

Jongheon Jeong*
KAIST
Deajeon, South Korea
jongheonj@kaist.ac.kr

Seojin Kim*
KAIST
Daejeon, South Korea
osikjs@kaist.ac.kr

Jinwoo Shin
KAIST
Daejeon, South Korea
jinwoos@kaist.ac.kr

Abstract

Any classifier can be "smoothed out" under Gaussian noise to build a new classifier that is provably robust to ℓ_2-adversarial perturbations, viz., by averaging its predictions over the noise via randomized smoothing. *In this paper, we propose a simple training method leveraging the fundamental trade-off between accuracy and (adversarial) robustness to obtain more robust smoothed classifiers, in particular, through a* sample-wise *control of robustness over the training samples. We make this control feasible by using "accuracy under Gaussian noise" as an easy-to-compute proxy of adversarial robustness for an input: specifically, we differentiate the training objective depending on this proxy to filter out samples that are unlikely to benefit from the worst-case (adversarial) objective. Our experiments show that the proposed method, despite its simplicity, consistently exhibits improved certified robustness upon state-of-the-art training methods. Somewhat surprisingly, we find these improvements persist even for other notions of robustness, e.g., to various types of common corruptions.*

1. Introduction

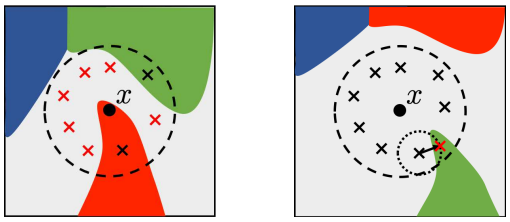
Despite these tremendous advances in *deep neural networks* for a variety of computer vision tasks towards artificial intelligence, the broad existence of *adversarial examples* [35] is still a significant aspect that reveals the gap between machine learning systems and humans: for a given input x (*e.g.*, an image) to a classifier f, say a neural network, f often permits a perturbation δ that completely flips the prediction $f(x+\delta)$, while δ is too small to change the semantic in x. In response to this vulnerability, there have been tremendous efforts in building *robust* neural network based classifiers against adversarial examples, either in forms of *empirical defenses* [1, 3, 36], which are largely based on *adversarial training* [26, 39, 41, 50, 51], or *certified defenses* [5, 40, 44, 48], depending on whether the robustness claim can be theoretically guaranteed or not.

Randomized smoothing [5, 21] is currently a prominent approach in the context of certified defense, thanks to its scalability to arbitrary neural network architectures while previous methods have been mostly limited in network sizes or require strong assumptions, *e.g.*, Lipschitz constraint, on their architectures. However, even with randomized smoothing, the *trade-off* between robustness and accuracy [37, 50] has been well evidenced, *i.e.*, increasing the robustness for a specific input can be at the expense of decreased accuracy for other inputs: *e.g.*, [50] has shown that the (empirical) robustness of a classifier can be further boosted in training by paying more expense in accuracy. A similar trend can be also observed with certified defenses, *e.g.*, randomized smoothing, as the clean accuracy of smoothed classifiers are usually less than those one can obtain from the standard training on the same architecture [5].

Contribution. In this paper, we develop a novel training method for randomized smoothing, coined *Confidence-Aware Training for Randomized Smoothing* (CAT-RS), which incorporates a *sample-wise* control of target robustness on-the-fly motivated by the accuracy-robustness trade-off in smoothed classifiers. Intuitively, a natural approach one can consider in response to the trade-off in robust training is to appropriately lower the robustness requirement for "hard-to-classify" samples while maintaining those for the remaining ("easier") samples: here, the challenges are (a) which samples should we choose for the control in training, and (b) how to control their target robustness. For both (a) and (b), the major difficultly stems from that evaluating adversarial robustness is computationally hard in practice. To implement this idea, we propose to use the sample-wise confidence of smoothed classifiers as an efficient proxy of the certified robustness, and defines two new losses, namely the *bottom-K* and *worst-case* Gaussian training, each of those targets different levels of confidence so that the overall training can prevent low-confidence samples from being enforced to increase their robustness.

We verify the effectiveness of our proposed method through an extensive comparison with existing state-of-the-art robust training methods for smoothed classifiers:

*Equal contribution



(a) Bottom-K loss (b) Worst-case loss

Figure 1. Illustration of the two proposed losses, *i.e.*, the (a) *bottom-K* and (b) *worst-case* Gaussian losses, respectively. Each $\times$ represents Gaussian noise around x. We aim to minimize the cross-entropy loss only for $\times$'s marked as red for each case.

it shows that CAT-RS consistently improves the previous state-of-the-art results on certified robustness, by (a) maximizing the robust radii of high-confidence samples while (b) reducing the risk of deteriorating the accuracy at low-confidence samples. We also find that CAT-RS also makes smoothed classifiers to generalize beyond adversarial robustness, from its significant gains in common corruption robustness: it confirms the importance of confidence information in adversarial training.

2. Preliminaries

Adversarial robustness. Consider an *i.i.d.* dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ from a certain distribution P, where $x \in \mathbb{R}^d$ and $y \in \mathcal{Y} := \{1, \cdots, K\}$, which forms a classification problem with K classes. Let $f : \mathbb{R}^d \to \mathcal{Y}$ be a (discrete) classifier. One can additionally consider a differentiable $F : \mathbb{R}^d \to \Delta^{K-1}$ to allow a gradient-based optimization assuming $f(x) := \arg\max_{k \in \mathcal{Y}} F_k(x)$, where Δ^{K-1} is probability simplex in $\mathbb{R}^K$. In the context of *adversarial robustness* (and for other notions of robustness as well), the *i.i.d.* assumption on the future samples does not hold anymore: instead, it additionally assumes that the samples can be *arbitrarily* perturbed up to a certain restriction, *e.g.*, a bounded ℓ_2-ball, and focuses on the *worst-case* performance over the perturbed samples. One possible way to quantify this scenario is to consider the *average minimum-distance* of adversarial perturbation [3, 4, 28], namely:

$$R(f; P) := \mathbb{E}_{(x,y)\sim P}\left[\min_{f(x')\neq y} ||x' - x||_2\right]. \quad (1)$$

Randomized smoothing. The essential challenge in achieving adversarial robustness in neural networks, however, stems from that directly evaluating (1) (and further optimizing it) is usually computationally infeasible. *Randomized smoothing* [5, 21] bypasses this difficulty by constructing a new classifier $\hat{f}$ from f instead of letting f to directly model the robustness: specifically, it transforms the base classifier f with a certain *smoothing measure*, where in this paper we focus on the case of Gaussian $\mathcal{N}(0, \sigma^2 I)$:

$$\hat{f}(x) := \arg\max_{c \in \mathcal{Y}} \mathbb{P}_{\delta\sim\mathcal{N}(0,\sigma^2 I)}\left(f(x+\delta) = c\right). \quad (2)$$

Then, the robustness of $\hat{f}$ at (x, y), namely $R(\hat{f}; x, y)$, can be explicitly lower-bounded in terms of the *certified radius* $\underline{R}(\hat{f}, x, y)$, *e.g.*, [5] showed that the following bound holds which is tight for ℓ_2-adversary:

$$R(\hat{f}; x, y) \geq \sigma \cdot \Phi^{-1}(p_f(x, y)) =: \underline{R}(\hat{f}, x, y) \quad (3)$$

$$\text{where} \quad p_f(x, y) := \mathbb{P}_\delta(f(x+\delta) = y), \quad (4)$$

provided that $\hat{f}(x) = y$, otherwise $R(\hat{f}; x, y) := 0$.[1] Here, we remark that the formula for certified radius (3) is essentially a function of p_f (4), which represents the *prediction confidence* of $\hat{f}$ at x, or equivalently, the *accuracy* of $f(x + \delta)$ over $\delta \sim \mathcal{N}(0, \sigma^2 I)$.

3. Confidence-aware Randomized Smoothing

We aim to develop a new training method to maximize the certified robustness of a smoothed classifier $\hat{f}$, considering the trade-off relationship between robustness and accuracy [50]: even though randomized smoothing can be applied for any classifier f, the actual robustness of $\hat{f}$ depends on how much f classifies well under presence of Gaussian noise, *i.e.*, by $p_f(x, y)$ defined in (4). A simple way to train f for a robust $\hat{f}$, therefore, is to minimize the standard cross-entropy loss $\mathbb{CE}$ with Gaussian noise as in [5]:

$$\min_F \mathbb{E}_{\substack{(x,y)\sim P \\ \delta\sim\mathcal{N}(0,\sigma^2 I)}}\left[\mathbb{CE}(F(x+\delta), y)\right]. \quad (5)$$

In this paper, we extend this basic form of training to incorporate a *confidence-aware* strategy to decide which noise samples $\delta_i \sim \mathcal{N}(0, \sigma^2 I)$ should be used for training. Ideally, given (4), one may wish to obtain a classifier f that achieves $p_f(x, y) \approx 1$ for every $(x, y) \sim P$ to maximize its certified robustness. In practice, however, such a case is highly unlikely, and there usually exists a sample x that $p_f(x, y)$ should be quite lower than 1 to maintain the discriminativity with other samples: in other words, these samples can be actually "beneficial" to be misclassified at some (hard) Gaussian noises, otherwise the classifier has to memorize the noises to correctly classify them. On the other hand, for the samples which can indeed achieve $p_f(x, y) \approx 1$, the current Gaussian training (5) may not be able to provide enough samples of δ_i for x throughout the training, as $p_f(x, y) \approx 1$ implies that $f(x + \delta)$ must be correctly classified "almost surely" for $\delta_i \sim \mathcal{N}(0, \sigma^2 I)$.

In these respects, we propose two different variants of Gaussian training (5) that address each of the possible cases,

[1] Φ denotes the *c.d.f.* of the standard normal distribution.

i.e., whether (a) $p_f(x,y) < 1$ or (b) $p_f(x,y) \approx 1$, namely with (a) *bottom-K* and (b) *worst-case* Gaussian training, respectively. During training, the method first estimates $p_f(x,y)$ for each sample by computing their accuracy over M random samples of $\delta \sim \mathcal{N}(0, \sigma^2 I)$, and applies different forms of loss depending on the value.

3.1. Bottom-K Loss for Low-confidence Samples

Consider a base classifier f and a training sample $(x,y) \in \mathcal{D}$, and suppose that $p_f(x,y) \ll 1$, *e.g.*, $\hat{f}$ has a low-confidence at x. Figure 1(a) visualizes this scenario: in this case, by definition of $p_f(x,y)$ in (4), $f(x+\delta)$ would be correctly classified to y only with probability p over $\delta \sim \mathcal{N}(0, \sigma^2 I)$, and this implies either (a) $x+\delta$ has not yet been adequately exposed to f during the training, or (b) $x+\delta$ may be indeed hard to be correctly classified for some δ, so that minimizing the loss at these noises could harm the generalization of $\hat{f}$. The design goal of our proposed *bottom-K Gaussian loss* is to modify the standard Gaussian training (5) to reduce the optimization burden from (b) while minimally retaining its ability to cover enough noise samples during training for (a).

We first assume M random *i.i.d.* samples of δ, say $\delta_1, \delta_2, \cdots, \delta_M \sim \mathcal{N}(0, \sigma^2 I)$. One can notice that the random variables $\mathbb{1}[f(x+\delta_i) = y]$'s are also *i.i.d.* each, which follows the Bernoulli distribution of probability $p_f(x,y)$. This means that, if the current $p_f(x,y)$ is the value one attempts to keep instead of further increasing it, the number of "correct" noise samples, namely $\sum_i \mathbb{1}[f(x+\delta_i) = y]$, would follow the *binomial distribution* $K \sim \mathrm{Bin}(M,p)$ - this motivates us to consider the following loss that only minimizes the K-*smallest* cross-entropy losses out of from M Gaussian samples around x:

$$L^{\texttt{low}} := \frac{1}{M} \sum_{i=1}^{K} \mathbb{CE}(F(x+\delta_{\pi(i)}), y), \tag{6}$$

where $K \sim \mathrm{Bin}(M, p_f(x,y))$. Here, $\pi(i)$ denotes the index with the i-th smallest loss value in the M samples.

3.2. Worst-case Loss for High-confidence Samples

Next, we focus on the case when $p_f(x,y) \approx 1$, *i.e.*, $\hat{f}$ has a high confidence at x, as illustrated in Figure 1(b). In contrast to the previous scenario in Section 3.1 (and Figure 1(a)), now the major drawback of Gaussian training (5) rather comes from the *rareness* of hard noises in training: considering that one can only present a limited number of noise samples to f throughout its training, naïvely minimizing (5) may not cover some "potentially hard" noise samples, and this would result in a significant harm in the final certified radius. The purpose of *worst-case* Gaussian training is to overcome this lack of samples via an *adversarial* search around each of the noise samples.

Specifically, for given M samples of Gaussian noise δ_i as considered in (6), namely $\delta_1, \cdots, \delta_M \sim \mathcal{N}(0, \sigma^2 I)$, we propose to modify (5) to find the *worst-case* noise δ^* (a) around an ℓ_2-ball for each noise as well as (b) among the M samples, and minimize the loss at δ^* instead of the average-case loss. To find such worst-case noise, our proposed loss optimizes a given δ_i to maximize the *consistency* of its prediction from a certain label assignment $\hat{y} \in \Delta^{K-1}$ per x:

$$L^{\texttt{high}} := \max_i \max_{\|\delta_i^* - \delta_i\|_2 \le \varepsilon} \mathrm{KL}(F(x+\delta_i^*), \hat{y}), \tag{7}$$

where $\mathrm{KL}(\cdot,\cdot)$ denotes the Kullback-Leibler divergence. This objective is motivated by [16] that the consistency of prediction across different Gaussian noise controls the trade-off between accuracy and robustness of smoothed classifiers. Notice from (7) that the objective is equivalent to the cross-entropy loss if $\hat{y}$ is assigned as (hard-labeled) y, while we observe having a soft-labeled $\hat{y}$ is beneficial in practice: its log-probability, where the consistency targets, can now be bounded so $F(x+\delta_i^*)$'s can also minimize their variance in the logit space.

There can be various ways to assign $\hat{y}$ for a given x: one of reasonable strategies, which we use in this paper as well, is to assign $\hat{y}$ by the *smoothed prediction* of another classifier $\bar{f}$, pre-trained on the same dataset $\mathcal{D}$ via Gaussian training (5) with some σ_0. This approach is straightforward to compute, and would (a) naturally reflect the sample-wise difficulty in classification under Gaussian noise, while (b) maintaining (most of) the label information given from y.

3.3. Overall Training Scheme

Given the two losses $L^{\texttt{low}}$ and $L^{\texttt{high}}$ defined in Section 3.1 and 3.2, respectively, we now define the full objective of our proposed *Confidence-Aware Training for Randomized Smoothing* (CAT-RS). Overall, in order to differentiate how to combine the two losses per sample basis, we use the smoothed confidence $p_f(x,y)$ (4) as the guiding proxy: specifically, we apply the worst-case loss of $L^{\texttt{high}}$ only for the samples where $p_f(x,y)$ is already high enough. In practice, we estimate $p_f(x,y)$ with the M noise samples *i.e.*, by $\hat{p}_f(x,y) := \frac{1}{M}\sum_{i=1}^{M} \mathbb{1}[f(x+\delta_i) = y]$. Then, we consider a simple and intuitive masking condition of "$K = M$" to activate $L^{\texttt{high}}$, where $K \sim \mathrm{Bin}(M, \hat{p}_f(x,y))$ is the random variable defined in (6) for $L^{\texttt{low}}$. The final loss becomes:

$$L^{\texttt{CAT-RS}} := L^{\texttt{low}} + \lambda \cdot \mathbb{1}[K = M] \cdot L^{\texttt{high}}, \tag{8}$$

where $\mathbb{1}[\cdot]$ is the indicator random variable, and $\lambda > 0$. The complete procedure of computing our proposed CAT-RS loss can be found in Algorithm 1 of Appendix A.

4. Experiments

We evaluate the effectiveness of our proposed training scheme based on various well-established image classifica-

Table 1. Comparison of ACR and approximate certified test accuracy (%) on CIFAR-10. For each column, we set our result bold-faced whenever the value improves the Gaussian baseline. We mark the highest and lowest values of certified accuracy at each radius in blue and red colors, respectively.

σ	Methods	ACR	0.00	0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00	2.25	2.50
0.25	Gaussian [5]	0.424	76.6	61.2	42.2	25.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Stability [23]	0.420	73.0	58.9	42.9	26.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	SmoothAdv [33]	0.544	73.4	65.6	57.0	47.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	MACER [46]	0.531	79.5	69.0	55.8	40.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Consistency [16]	0.552	75.8	67.6	58.1	46.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	SmoothMix [15]	0.553	77.1	67.9	57.9	46.7	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	CAT-RS (Ours)	**0.562**	76.3	**68.1**	**58.8**	**48.2**	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.50	Gaussian [5]	0.525	65.7	54.9	42.8	32.5	22.0	14.1	8.3	3.9	0.0	0.0	0.0
	Stability [23]	0.531	62.1	52.6	42.7	33.3	23.8	16.1	9.8	4.7	0.0	0.0	0.0
	SmoothAdv [33]	0.684	65.3	57.8	49.9	41.7	33.7	26.0	19.5	12.9	0.0	0.0	0.0
	MACER [46]	0.691	64.2	57.5	49.9	42.3	34.8	27.6	20.2	12.6	0.0	0.0	0.0
	Consistency [16]	0.720	64.3	57.5	50.6	43.2	36.2	29.5	22.8	16.1	0.0	0.0	0.0
	SmoothMix [15]	0.737	61.8	55.9	49.5	43.3	37.2	31.7	25.7	19.8	0.0	0.0	0.0
	CAT-RS (Ours)	**0.757**	62.3	**56.8**	**50.5**	**44.6**	**38.5**	**32.7**	**27.1**	**20.6**	0.0	0.0	0.0
1.00	Gaussian [5]	0.511	47.1	40.9	33.8	27.7	22.1	17.2	13.3	9.7	6.6	4.3	2.7
	Stability [23]	0.514	43.0	37.8	32.5	27.5	23.1	18.8	14.7	11.0	7.7	5.2	3.1
	SmoothAdv [33]	0.790	43.7	40.3	36.9	33.8	30.5	27.0	24.0	21.4	18.4	15.9	13.4
	MACER [46]	0.744	41.4	38.5	35.2	32.3	29.3	26.4	23.4	20.2	17.4	14.5	12.1
	Consistency [16]	0.756	46.3	42.2	38.1	34.3	30.0	26.3	22.9	19.7	16.6	13.8	11.3
	SmoothMix [15]	0.773	45.1	41.5	37.5	33.8	30.2	26.7	23.4	20.2	17.2	14.7	12.1
	CAT-RS (Ours)	**0.815**	43.2	40.2	**37.2**	**34.3**	**31.0**	**28.1**	**24.9**	**22.0**	**19.3**	**16.8**	**14.2**

tion benchmarks, including MNIST [20], Fashion-MNIST [43], CIFAR-10/100, and ImageNet [18].[2] For a fair comparison, we follow the standard protocol and training setup of the previous works [5,15,16,46]:[3] specifically, we use (a) the *average certified radius* (ACR) [46] and (b) the *approximate certified test accuracy* at r as the major performance metrics throughout experiments.[4]

4.1. Certified Adversarial Robustness

We compare the certified robustness of the smoothed classifiers trained on CIFAR-10 in Table 3, considering three different smoothing factors $\sigma \in \{0.25, 0.5, 1.0\}$.[5] For the baselines, we report best-performing configurations for each σ in terms of ACR among reported in previous works, so that the hyperparameters of the same method can vary over σ (the details can be found in Appendix C.5). Overall, CAT-RS achieves a significant improvement of ACR compared to the baselines. In case of $\sigma = 0.25$ and $\sigma = 0.5$, CAT-RS clearly offers a better trade-off between the clean accuracy and robustness compared to other baselines. Especially, CAT-RS achieves higher approximate certified accuracy for all radii compared to SmoothMix in case of $\sigma = 0.5$. For $\sigma = 1.0$, the ACR of our method significantly surpasses the previous best model, SmoothMix, by $0.773 \rightarrow 0.815$. Remarkably, the improvement from CAT-RS is most evident in $\sigma = 1.0$, suggesting the effectiveness of confidence-aware training in adversarial robustness.

[2]Results on more datasets, *viz.*, MNIST, Fashion-MNIST, CIFAR-100, and ImageNet can be found in Appendix D.

[3]The full details, *e.g.*, training setups, baselines, evaluation metrics, and hyperparameters, can be found in Appendix C.

[4]We also perform an ablation study in Appendix F, showing that, *e.g.*, the major hyperpramter λ (8) can effectively balance the accuracy-robustness trade-off, which is favorable in practical uses.

[5]Figure 3 in Appendix also plots the certified accuracy over r.

Table 2. Comparison of *average certified radius* (ACR) on CIFAR-10-C. We report the average across five different corruption severities. We set the highest values bold-faced for each row. We set the runner-up values underlined.

Type	Gaussian [5]	Stability [23]	SmoothAdv [33]	MACER [46]	Consistency [16]	SmoothMix [15]	**CAT-RS (Ours)**
Gaussian	0.412	0.348	0.506	0.473	0.505	0.513	**0.544**
Shot	0.414	0.350	0.503	0.472	0.503	0.508	**0.542**
Impulse	0.389	0.322	0.495	0.452	0.492	0.499	**0.530**
Defocus	0.372	0.329	0.480	0.442	0.482	0.489	**0.512**
Glass	0.343	0.291	0.473	0.415	0.472	0.483	**0.505**
Motion	0.352	0.314	0.458	0.417	0.465	0.474	**0.492**
Zoom	0.346	0.315	0.468	0.420	0.462	0.476	**0.501**
Snow	0.346	0.325	0.452	0.417	0.448	0.438	**0.487**
Frost	0.298	0.298	**0.434**	0.377	0.401	0.403	**0.434**
Fog	0.197	0.153	0.279	0.266	0.277	0.262	**0.293**
Bright	0.378	0.366	0.487	0.451	0.489	0.478	**0.524**
Constrast	0.146	0.131	**0.228**	0.195	0.213	0.202	**0.228**
Elastic	0.331	0.290	0.441	0.405	0.445	0.447	**0.464**
Pixel	0.404	0.350	0.500	0.465	0.500	0.509	**0.538**
JPEG	0.413	0.354	0.504	0.470	0.502	0.504	**0.537**
mACR	0.343	0.302	0.447	0.409	0.444	0.446	**0.475**

4.2. Corruption Robustness

We also examine the performance of our training method on CIFAR-10-C [13], a collection of 75 replicas of the CIFAR-10 test dataset, which consists of 15 different types of common corruptions (*e.g.*, fog, snow, etc.), each of which contains 5 levels of corruption severities.[6] For a given smoothed classifier trained on ("clean") CIFAR-10, we report ACR for each corruption type of CIFAR-10-C after averaging the values over five severity levels, as well as their means over the types, *i.e.*, as the mean-ACR (mACR).[7] Here, we uniformly subsample each corrupted dataset with size 100, *i.e.*, to have 7,500 in total, and use $\sigma = 0.25$.

Table 2 summarizes the results. Overall, we observe that CAT-RS consistently achieves the best ACRs on all the corruption types, thus also in mACR. In particular, we find CAT-RS can better maintain the ("clean") ACR given in Table 1 ($\sigma = 0.25$) under corruptions compared to other methods, as shown in the reduced overall gaps in ACR. In other words, CAT-RS can improve smoothed classifiers to generalize better on unseen corruptions, at the same time maintaining the robustness for such inputs. It is remarkable that the observed gains are not from any prior knowledge about multiple corruption [12,14] (except for Gaussian noise), but from a better training method. Given the limited gains from other baseline methods on CIFAR-10-C, we attribute that the *sample-dependent calibration* of training objective, a unique aspect of CAT-RS compared to prior arts, is important to explain the effectiveness of CAT-RS on out-of-distribution generalization: *e.g.*, although SmoothAdv also adopts adversarial search in training similarly to CAT-RS, it could not improve mAcc on CIFAR-10-C from Gaussian.

[6]Additional results on MNIST-C [29] can be also found in Appendix I.

[7]We also report the certified accuracy at $r = 0.0$ and the mean-accuracy (mAcc) with more detailed results in Appendix H, showing that CAT-RS also achieves the best mAcc compared to other methods.

References

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

[2] Mislav Balunovic and Martin Vechev. Adversarial training and provable defenses: Bridging the gap. In *International Conference on Learning Representations*, 2020. 8

[3] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, and Aleksander Madry. On evaluating adversarial robustness. *arXiv preprint arXiv:1902.06705*, 2019. 1, 2

[4] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, pages 39–57. IEEE, 2017. 2

[5] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1310–1320, Long Beach, California, USA, 09–15 Jun 2019. PMLR. 1, 2, 4, 8, 9, 11, 12, 13, 16, 17, 18, 19, 20, 21

[6] Francesco Croce, Maksym Andriushchenko, and Matthias Hein. Provable robustness of ReLU networks via maximization of linear regions. In Kamalika Chaudhuri and Masashi Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 2057–2066. PMLR, 16–18 Apr 2019. 8

[7] Francesco Croce and Matthias Hein. Provable robustness against all adversarial l_p-perturbations for $p \geq 1$. In *International Conference on Learning Representations*, 2020. 8

[8] Marc Fischer, Maximilian Baader, and Martin Vechev. Scalable certified segmentation via randomized smoothing. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 3340–3351. PMLR, 18–24 Jul 2021. 8

[9] 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*, pages 3–18. IEEE, 2018. 8

[10] Sven Gowal, Krishnamurthy Dj Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Mann, and Pushmeet Kohli. Scalable verified training for provably robust image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4842–4851, 2019. 8

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 8, 13, 14

[12] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8340–8349, October 2021. 4

[13] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. 4

[14] Dan Hendrycks*, Norman Mu*, Ekin Dogus Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. AugMix: A simple method to improve robustness and uncertainty under data shift. In *International Conference on Learning Representations*, 2020. 4

[15] Jongheon Jeong, Sejun Park, Minkyu Kim, Heung-Chang Lee, Do-Guk Kim, and Jinwoo Shin. SmoothMix: Training confidence-calibrated smoothed classifiers for certified robustness. *Advances in Neural Information Processing Systems*, 34:30153–30168, 2021. 4, 8, 9, 10, 11, 12, 13, 15, 16, 17, 18, 19, 20, 21

[16] Jongheon Jeong and Jinwoo Shin. Consistency regularization for certified robustness of smoothed classifiers. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 10558–10570. Curran Associates, Inc., 2020. 3, 4, 8, 9, 10, 11, 12, 13, 15, 16, 17, 18, 19, 20, 21

[17] Jinyuan Jia, Xiaoyu Cao, Binghui Wang, and Neil Zhenqiang Gong. Certified robustness for top-k predictions against adversarial perturbations via randomized smoothing. In *International Conference on Learning Representations*, 2020. 8

[18] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, Department of Computer Science, University of Toronto, 2009. 4, 9, 12

[19] Aounon Kumar, Alexander Levine, Soheil Feizi, and Tom Goldstein. Certifying confidence via randomized smoothing. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 5165–5177. Curran Associates, Inc., 2020. 8

[20] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, Nov 1998. 4, 8, 9, 21

[21] 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)*, pages 656–672. IEEE, 2019. 1, 2

[22] Guang-He Lee, Yang Yuan, Shiyu Chang, and Tommi Jaakkola. Tight certificates of adversarial robustness for randomly smoothed classifiers. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 8

[23] Bai Li, Changyou Chen, Wenlin Wang, and Lawrence Carin. Certified adversarial robustness with additive noise. In *Advances in Neural Information Processing Systems 32*, pages

9464–9474. Curran Associates, Inc., 2019. 4, 9, 10, 11, 12, 13, 16, 17, 18, 19, 20, 21

[24] Linyi Li, Xiangyu Qi, Tao Xie, and Bo Li. SoK: Certified robustness for deep neural networks. *arXiv preprint arXiv:2009.04131*, 2020. 8

[25] Linyi Li, Maurice Weber, Xiaojun Xu, Luka Rimanic, Bhavya Kailkhura, Tao Xie, Ce Zhang, and Bo Li. TSS: Transformation-specific smoothing for robustness certification. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, page 535–557, New York, NY, USA, 2021. Association for Computing Machinery. 8

[26] 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. 1, 8, 9

[27] Matthew Mirman, Timon Gehr, and Martin Vechev. Differentiable abstract interpretation for provably robust neural networks. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3578–3586, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. 8

[28] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. DeepFool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2574–2582, 2016. 2

[29] Norman Mu and Justin Gilmer. MNIST-C: A robustness benchmark for computer vision, 2019. 4, 21

[30] Elan Rosenfeld, Ezra Winston, Pradeep Ravikumar, and Zico Kolter. Certified robustness to label-flipping attacks via randomized smoothing. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8230–8241. PMLR, 13–18 Jul 2020. 8

[31] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 8, 9, 13

[32] Hadi Salman, Saachi Jain, Eric Wong, and Aleksander Madry. Certified patch robustness via smoothed vision transformers, 2021. 8

[33] Hadi Salman, Jerry Li, Ilya Razenshteyn, Pengchuan Zhang, Huan Zhang, Sebastien Bubeck, and Greg Yang. Provably robust deep learning via adversarially trained smoothed classifiers. In *Advances in Neural Information Processing Systems 32*, pages 11289–11300. Curran Associates, Inc., 2019. 4, 8, 9, 10, 11, 12, 13, 16, 17, 18, 19, 20, 21

[34] Hadi Salman, Mingjie Sun, Greg Yang, Ashish Kapoor, and J. Zico Kolter. Denoised smoothing: A provable defense for pretrained classifiers. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21945–21957. Curran Associates, Inc., 2020. 8

[35] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014. 1

[36] Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. In *Advances in Neural Information Processing Systems*, volume 33, 2020. 1

[37] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019. 1

[38] Binghui Wang, Jinyuan Jia, Xiaoyu Cao, and Neil Zhenqiang Gong. Certified robustness of graph neural networks against adversarial structural perturbation. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, page 1645–1653, New York, NY, USA, 2021. Association for Computing Machinery. 8

[39] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations*, 2020. 1, 8

[40] Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5286–5295, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. 1, 8

[41] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 2958–2969. Curran Associates, Inc., 2020. 1

[42] Fan Wu, Linyi Li, Zijian Huang, Yevgeniy Vorobeychik, Ding Zhao, and Bo Li. CROP: Certifying robust policies for reinforcement learning through functional smoothing. In *International Conference on Learning Representations*, 2022. 8

[43] Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017. 4, 9, 11

[44] Kai Y. Xiao, Vincent Tjeng, Nur Muhammad (Mahi) Shafiullah, and Aleksander Madry. Training for faster adversarial robustness verification via inducing ReLU stability. In *International Conference on Learning Representations*, 2019. 1, 8

[45] Greg Yang, Tony Duan, J. Edward Hu, Hadi Salman, Ilya Razenshteyn, and Jerry Li. Randomized smoothing of all shapes and sizes. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10693–10705. PMLR, 13–18 Jul 2020. 8

[46] Runtian Zhai, Chen Dan, Di He, Huan Zhang, Boqing Gong, Pradeep Ravikumar, Cho-Jui Hsieh, and Liwei Wang.

MACER: Attack-free and scalable robust training via maximizing certified radius. In *International Conference on Learning Representations*, 2020. 4, 8, 9, 10, 11, 12, 13, 16, 17, 18, 19, 20, 21

[47] Dinghuai Zhang, Mao Ye, Chengyue Gong, Zhanxing Zhu, and Qiang Liu. Black-box certification with randomized smoothing: A functional optimization based framework. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 2316–2326. Curran Associates, Inc., 2020. 8

[48] Huan Zhang, Hongge Chen, Chaowei Xiao, Sven Gowal, Robert Stanforth, Bo Li, Duane Boning, and Cho-Jui Hsieh. Towards stable and efficient training of verifiably robust neural networks. In *International Conference on Learning Representations*, 2020. 1, 8

[49] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018. 9

[50] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7472–7482, Long Beach, California, USA, 09–15 Jun 2019. PMLR. 1, 2, 14

[51] Jingfeng Zhang, Xilie Xu, Bo Han, Gang Niu, Lizhen Cui, Masashi Sugiyama, and Mohan Kankanhalli. Attacks which do not kill training make adversarial learning stronger. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11278–11287. PMLR, 13–18 Jul 2020. 1

[52] Jingfeng Zhang, Jianing Zhu, Gang Niu, Bo Han, Masashi Sugiyama, and Mohan Kankanhalli. Geometry-aware instance-reweighted adversarial training. In *International Conference on Learning Representations*, 2021. 8