Average Certified Radius is a Poor Metric for Randomized Smoothing

Chenhao Sun* | Yuhao Mao* | Mark Niklas Müller | Martin Vechev

ETH zürich · SRILAB · logic*

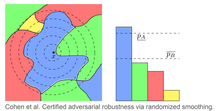
We theoretically analyze the weakness of a widely used metric (ACR) for Randomized Smoothing

We exploit the weakness of ACR and achieve a SOTA ACR

Stop using ACR and try ECDF!

Background

Randomized smoothing



Cohen et al. Certified adversarial robustness via randomized smoothing.

$$g(x) = \arg\max_{c \in \mathcal{Y}} \mathbb{P}(f(x+\epsilon) = c)$$

Smooth model | Base model | Gaussian noise

Certified radius

$g(x)$ is adversarially robust in an ℓ_2 norm neighborhood of x with the radius $R(x)$

Average certified radius

$$\text{ACR} := \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} R(x) \mathbf{1}(g(x) = y)$$

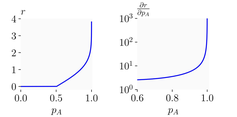
Weakness of ACR

A trivial classifier can achieve infinite ACR

$$\text{ACR} := \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} R(x, p_A) \mathbf{1}(g(x) = y)$$

As large as you want with enough budget

Easy samples contribute much more to the ACR



p_A: 0.99 → 0.999 r: 2.3 → 3.0

p_A: 0.0 → 0.76 r: 0.0 → 0.69

Selection bias in RS training algorithms

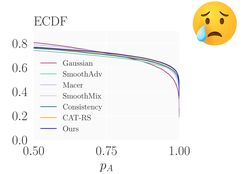
A Poor Metric!

Previous methods implicitly focus on easy data and sacrifice hard data

Method	ACR	easy	hard	easy / hard
Gaussian	0.56	10.10	22.67	0.45
SmoothAdv	0.68	5.60	5.62	1.00
Consistency	0.72	14.99	19.32	0.78
SmoothMix	0.74	11.72	11.79	0.99
CAT-RS	0.76	30.45	7.12	4.28

Table 2: The average model parameters gradient l_2 norm of easy ($p_A > 0.5$) and hard ($p_A < 0.5$) samples for models trained with different algorithms and $\sigma = 0.5$, along with their relative magnitude (*easy / hard*). The corresponding ACR is provided for reference.

SOTA methods tend to assign greater emphasis to easy samples than Gaussian training.

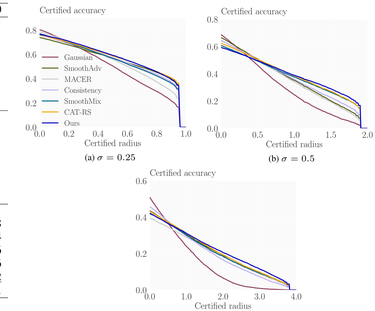


SOTA methods gain ACR due to the improvement on easy data; hard data are consistently under-represented

Main Results

Certification results on CIFAR-10

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	0.486	**81.3**	66.7	50.0	32.4	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	MACER	0.529	78.7	68.3	55.9	40.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	SmoothAdv	0.544	73.4	65.6	57.0	47.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Consistency	0.547	75.8	67.4	57.5	46.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	SmoothMix	0.543	77.1	67.6	56.8	45.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	CAT-RS	0.562	76.3	68.1	58.8	48.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Ours	**0.564**	76.6	**69.1**	**59.3**	**48.3**	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.5	Gaussian	0.562	**68.7**	57.6	45.7	34.0	23.7	15.9	9.4	4.8	0.0	0.0	0.0
	MACER	0.680	64.7	57.4	49.5	42.1	34.0	26.4	19.2	12.0	0.0	0.0	0.0
	SmoothAdv	0.684	65.3	**57.8**	49.9	41.7	33.7	26.0	19.5	12.9	0.0	0.0	0.0
	Consistency	0.716	64.1	57.6	50.3	42.9	35.9	29.1	22.6	16.0	0.0	0.0	0.0
	SmoothMix	0.738	60.6	55.2	49.3	43.3	37.6	32.1	26.4	20.5	0.0	0.0	0.0
	CAT-RS	0.757	62.3	56.8	**50.5**	**44.6**	38.5	32.7	27.1	20.6	0.0	0.0	0.0
	Ours	**0.760**	59.3	54.8	49.6	44.4	**38.9**	**34.1**	**29.0**	**23.0**	0.0	0.0	0.0
1.0	Gaussian	0.534	**51.5**	**44.1**	36.5	29.4	23.8	18.2	13.1	9.2	6.0	3.8	2.3
	MACER	0.760	39.5	36.9	34.6	31.7	28.9	26.4	23.8	21.1	18.6	16.0	13.8
	SmoothAdv	0.790	43.7	40.3	36.9	33.8	30.5	27.0	24.0	21.4	18.4	15.9	13.4
	Consistency	0.757	45.7	42.0	**37.8**	33.7	30.0	26.3	22.9	19.6	16.6	13.9	11.6
	SmoothMix	0.788	42.4	39.4	36.7	33.4	30.0	26.8	23.9	20.8	18.6	15.9	13.6
	CAT-RS	0.815	43.2	40.2	37.2	**34.3**	31.0	28.1	24.9	22.0	19.3	16.8	14.2
	Ours	**0.844**	42.0	39.4	36.5	33.9	**31.1**	**28.4**	**25.6**	**23.1**	**20.6**	**18.3**	**16.1**



(a) $\sigma = 0.25$ (b) $\sigma = 0.5$ (c) $\sigma = 1.0$

Certification results on ImageNet

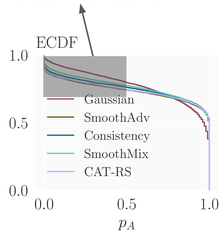
σ	Methods	ACR	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5
0.25	Gaussian	0.476	**66.7**	49.4	0.0	0.0	0.0	0.0	0.0	0.0
	SmoothAdv	**0.532**	65.3	55.5	0.0	0.0	0.0	0.0	0.0	0.0
	Ours	0.529	64.2	**55.8**	0.0	0.0	0.0	0.0	0.0	0.0
0.5	Gaussian	0.733	**57.2**	45.8	37.2	28.6	0.0	0.0	0.0	0.0
	SmoothAdv	0.824	53.6	49.4	43.3	36.8	0.0	0.0	0.0	0.0
	Consistency	0.822	55.0	**50.2**	**44.0**	34.8	0.0	0.0	0.0	0.0
	SmoothMix	**0.846**	54.6	50.0	43.4	**37.8**	0.0	0.0	0.0	0.0
	Ours	0.842	55.5	48.6	43.8	37.6	0.0	0.0	0.0	0.0
1.0	Gaussian	0.875	**43.6**	37.8	32.6	26.0	19.4	14.8	12.2	9.0
	SmoothAdv	1.040	40.3	37.0	34.0	30.0	26.9	**24.6**	19.7	15.2
	Consistency	0.982	41.6	37.4	32.6	28.0	24.2	21.0	17.4	14.2
	SmoothMix	1.047	39.6	37.2	33.6	30.4	26.2	24.0	**20.4**	**17.0**
	CAT-RS	**1.071**	**43.6**	**38.2**	**35.2**	30.8	26.8	24.0	20.2	**17.0**
	Ours	1.042	41.0	37.4	33.6	**31.0**	**27.0**	23.2	19.4	15.8

Ablation study ($\sigma = 0.25$)

discard	dataset weight	adverserial	ACR	0.00	0.25	0.50	0.75
	Gaussian		0.486	81.3	66.7	50.0	32.4
✔			0.515	81.2	69.3	53.7	36.8
	✔		0.512	81.3	69.4	53.3	36.3
		✔	0.537	76.7	66.7	55.6	44.3
✔	✔		0.523	81.1	69.7	54.6	38.3
✔		✔	0.550	77.4	68.5	57.7	45.4
	✔	✔	0.554	75.0	67.1	58.1	48.1
✔	✔	✔	**0.564**	76.6	69.1	59.3	48.3

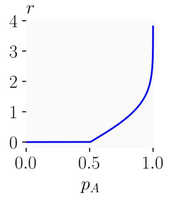
Replicating ACR Progress WITHOUT Training for General Robustness

Does not affect ACR



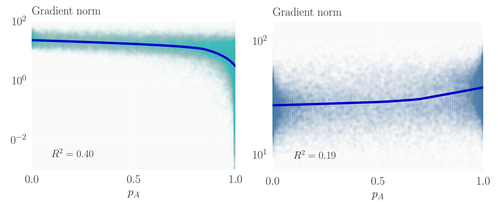
Discard hard data

Disproportionate contributions to ACR



Reweight training data

No effective optimization for easy data



(a) Before adaptive attack (b) After adaptive attack

Effective noise search

SOTA ACR!

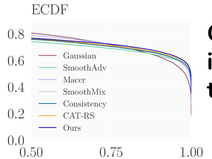
Better Metrics

Best certified accuracy across σ at each radius

- At the radius of interest, one may identify the most effective method and its associated noise level to optimize performance on downstream tasks.

Empirical distribution of p_A

- One method is better than another only when it has higher ECDF for all $p_A \geq 0.5$.
- Easy to convert ECDF to the certified accuracy for every radius and certification budget.
- Convenient to convert an existing certified accuracy-radius curve to the ECDF of p_A.



(a) $\sigma = 0.25$

Gaussian training is still the best in the low p_A region!