Confidence-aware Training of Smoothed Classifiers for Certified Robustness

Jongheon Jeong* (KAIST EE; jongheonj@kaist.ac.kr), Seojin Kim* (KAIST AI), Jinwoo Shin (KAIST AI)

CAT-RS = A new, state-of-the-art robust training for randomized smoothing

- **Motivation**: For some inputs, achieving high robust radius is fundamentally challenging
- Appropriately calibrating (smoothed) confidences impacts the certified robustness

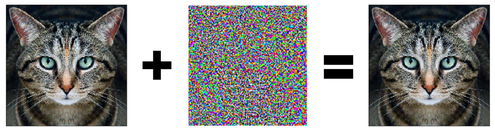
Background

Adversarial examples in deep learning are easy-to-find and pervasive

- A small, worst-case input noise that affects the output prediction

Goal: $f(\mathbf{x}) = f(\mathbf{x}+\boldsymbol{\delta}), \quad \boxed{\forall \boldsymbol{\delta}} : \|\boldsymbol{\delta}\|_2 \leq \varepsilon$

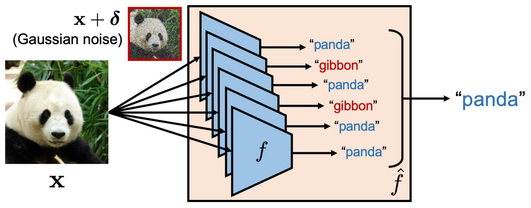
a classifier — The hardest part



Randomized smoothing (RS) builds a new classifier from any base classifier f

$$\hat{f}(\mathbf{x}) := \arg\max_{k \in \mathcal{Y}} \{ \mathbb{P}_{\boldsymbol{\delta} \sim \mathcal{N}(0, \sigma^2 I)} (f(\mathbf{x}+\boldsymbol{\delta}) = k) \}$$

Gaussian noise



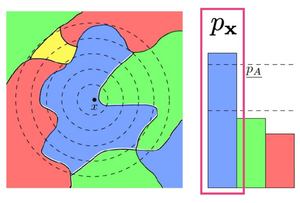
- **Idea**: Define a **smoothed classifier** by averaging f over Gaussian noise
- **Cohen et al. (2019)**: A provable guarantee on adversarial robustness of $\hat{f}$

Theorem Let $p_{\mathbf{x}} := \max_k \{ \mathbb{P}_{\boldsymbol{\delta}} (f(\mathbf{x}+\boldsymbol{\delta}) = k) \}$.
The ℓ_2-robust radius at $\hat{f}(\mathbf{x})$ is **lower-bounded** by:

$$R(\hat{f}; \mathbf{x}) := \min_{\hat{f}(\mathbf{x}+\boldsymbol{\delta}) \neq \hat{f}(\mathbf{x})} \|\boldsymbol{\delta}\|_2 \geq \sigma \cdot \Phi^{-1}(p_{\mathbf{x}})$$

Gaussian CDF

Remark: The higher $p_{\mathbf{x}}$, the better robustness at $\hat{f}(\mathbf{x})$



Randomized smoothing (RS) introduces a new problem:

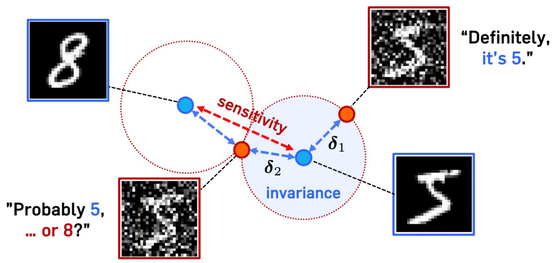
(AT) "How to train f to maximize the robustness of f?"

⇨ (RS) "How to train f to maximize the robustness of $\hat{f}$?"

Confidence-aware Randomized Smooting

Motivation: Invariance to noise is often at a cost of model sensitivity [Tramer et al., 2020]

⇒ For some cases, achieving "high RS confidence" is challenging even for humans



Case 1: For some noise samples $\boldsymbol{\delta}$, $\mathbf{x}+\boldsymbol{\delta}$ is fundamentally "hard-to-classify"

Bottom-K Gaussian objective: $\boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta}_M \sim \mathcal{N}(0, \sigma^2 I)$

$$L^{\text{low}} := \frac{1}{M} \sum_{i=1}^{K} \mathbb{CE}(F(\mathbf{x}+\boldsymbol{\delta}_{\pi(i)}), y),$$

Sorted index (ascending) w.r.t. the loss values

where $K \sim \text{Bin}(M, \hat{p}_{\mathbf{x}})$.

Binomial distribution — Any approximation of $p_{\mathbf{x}}$

Case 2: The standard Gaussian training may not fully cover "potentially hard" noises

Worst-case Gaussian objective: $\boldsymbol{\delta}_1, \cdots, \boldsymbol{\delta}_M \sim \mathcal{N}(0, \sigma^2 I)$

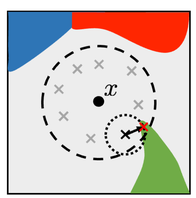
Kullback-Leibler divergence

$$L^{\text{high}} := \max_i \ \text{KL}(F(\mathbf{x}+\boldsymbol{\delta}_i^*) \parallel \hat{y}),$$

"Select only the true worst-case"

where $\boldsymbol{\delta}_i^* := \arg\max_{\|\boldsymbol{\delta}_i^* - \boldsymbol{\delta}_i\|_2 \leq \varepsilon} \ \text{KL}(F(\mathbf{x}+\boldsymbol{\delta}_i^*) \parallel \hat{y})$.

Adversarial search on noise — A soft-label assignment : e.g., $\hat{y} := \frac{1}{M} \sum_i F(\mathbf{x}+\boldsymbol{\delta}_i)$



CAT-RS differently applies L^{low} and L^{high} sample-wise using M Gaussian noises

🤔 How to decide which objective to use per sample?

💡 A simple masking condition "$K = M$": i.e., when L^{low} covers the full noise samples

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

Experiments

CAT-RS consistently improves (certified) adversarial robustness

- Better accuracy-robustness trade-off compared to SOTAs, e.g., SmoothMix

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	0.424	76.6	61.2	42.2	25.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Stability	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	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	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	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	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	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	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	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	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	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	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	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	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	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	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	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	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**

Comparison of ACR and certified accuracy on CIFAR-10 (ResNet-110, $\sigma \in \{0.25, 0.5, 1.0\}$)

- CAT-RS also successfully scales up to certify on large-scale ImageNet

Methods	ACR	0.0	0.5	1.0	1.5	2.0	2.5	3.0	3.5
Gaussian (Cohen et al. 2019)	0.875	**44**	**38**	33	26	19	15	12	9
Consistency (Jeong and Shin 2020)	0.982	41	37	32	28	24	21	17	14
SmoothAdv (Salman et al. 2019)	1.003	40	37	34	30	**27**	**25**	**20**	15
SmoothMix (Jeong et al. 2021)	1.047	40	37	34	30	26	24	**20**	**17**
CAT-RS (Jeong et al. 2022)	**1.071**	**44**	**38**	**35**	**31**	**27**	24	**20**	**17**

Results on ImageNet (ResNet-50, $\sigma = 0.5$)

- CAT-RS can improve RS to further generalize on unseen corruptions

Type	Gaussian	Stability	SmoothAdv	MACER	Consistency	SmoothMix	**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**

Type	Gaussian	Stability	SmoothAdv	MACER	Consistency	SmoothMix	**CAT-RS (Ours)**
Clean	76.6	73.0	73.4	**79.5**	75.8	77.1	76.3
Gaussian	70.8	64.6	70.2	72.6	69.8	73.4	**76.8**
Shot	70.0	65.6	68.4	72.8	69.6	72.6	**76.6**
Impulse	70.2	61.6	69.0	74.0	70.4	73.6	**75.6**
Defocus	64.8	65.4	68.4	71.2	69.2	70.6	**74.2**
Glass	65.2	62.0	68.6	71.6	69.0	72.0	**72.8**
Motion	66.2	62.4	67.2	**72.2**	70.8	69.6	71.6
Zoom	65.2	64.2	65.6	70.6	68.4	71.4	**75.4**
Snow	67.0	64.6	64.0	70.8	67.0	69.2	**71.4**
Frost	65.6	63.0	64.0	69.0	66.8	**70.2**	67.8
Fog	52.4	38.8	45.4	**53.8**	49.2	50.4	51.4
Bright	71.0	70.6	67.6	73.8	73.2	73.8	**76.4**
Constrast	39.4	30.0	34.8	**42.8**	35.6	36.4	37.8
Elastic	64.4	63.4	64.6	71.0	66.4	69.8	**71.4**
Pixel	66.4	67.6	68.6	74.4	69.8	69.8	**76.2**
JPEG	67.8	66.8	68.6	70.8	68.4	70.8	**76.2**
mAcc	64.4	60.7	63.7	68.8	65.6	67.7	**70.1**

Results on CIFAR-10-C (ResNet-110, $\sigma = 0.25$)