ICML | 2022

Thirty-ninth International Conference on Machine Learning

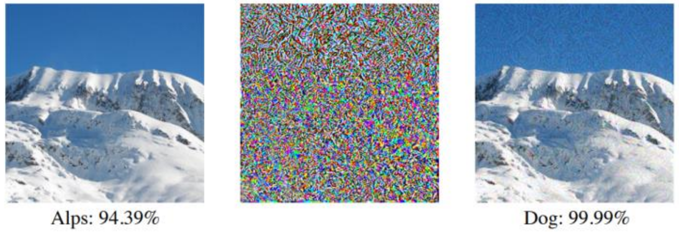
GSmooth: Certified Robustness against Semantic Transformations via Generalized Randomized Smoothing

Zhongkai Hao, Chengyang Ying, Yinpeng Dong, Hang Su, Jun Zhu, Jian Song,

Department of Computer Science and Technology, Tsinghua University

Motivation

- Certified defense provides a promising method for adaptive attacks



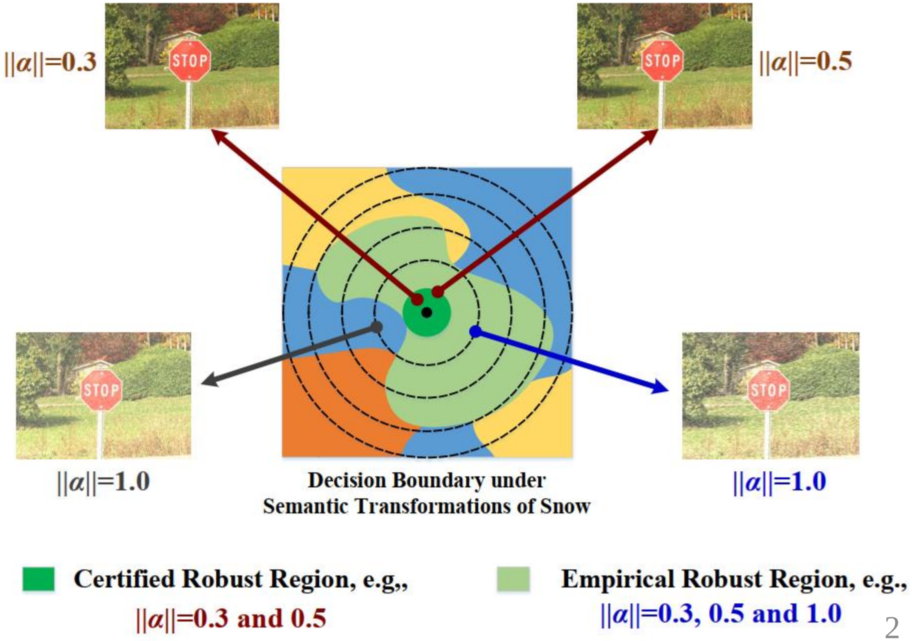
- Certified defense calculates a radius that the worst case classifier is still right

$$a = \arg\max_{y \in \mathcal{Y}} G(x)_y \qquad and \qquad b = \arg\max_{y \in \mathcal{Y} \setminus \{y^*\}} G(x)_y$$

- Then for the corrupted x' we have

$$\arg\max_{y \in \mathcal{Y}} G(x')_y = a$$

- Can we defense semantic transformations certifiably ?



GSmooth: Generalized Randomized Smoothing

- We define the smoothed classifier for a soft classifier f and semantic transformation τ as

$$G(x) = \mathbb{E}_{\theta \sim g(\cdot)}[f(\tau(\theta, x))],$$

- Restate the certified bound for resolvable transformations, where $\gamma(\theta, \xi)$ is the new parameter under composition

Theorem 1. *Let $f(x)$ be any classifier and $G(x)$ be the smoothed classifier defined in Eq. (1). If there exists a function $M(\cdot,\cdot) : P \times P \to \mathbb{R}$, the transformation $\tau(\cdot,\cdot)$ satisfies*

$$\frac{\partial \gamma(\theta, \xi)}{\partial \xi} = \frac{\partial \gamma(\theta, \xi)}{\partial \theta} M(\theta, \xi),$$

and there exist two constants $\underline{p_A}$, $\overline{p_B}$ satisfying

$$G(x)_A \geqslant \underline{p_A} \geqslant \overline{p_B} \geqslant G(x)_B,$$

then $y_A = \arg\max_{i \in \mathcal{Y}} G(\tau(\xi, x))_i$ holds for any $\|\xi\| \leqslant R$ where

$$R = \frac{1}{2M^*} \int_{\overline{p_B}}^{\underline{p_A}} \frac{1}{\Phi(p)} dp, \tag{3}$$

and $M^* = max_{\xi, \theta \in P} ||M(\xi, \theta)||$.

GSmooth: Generalized Randomized Smoothing

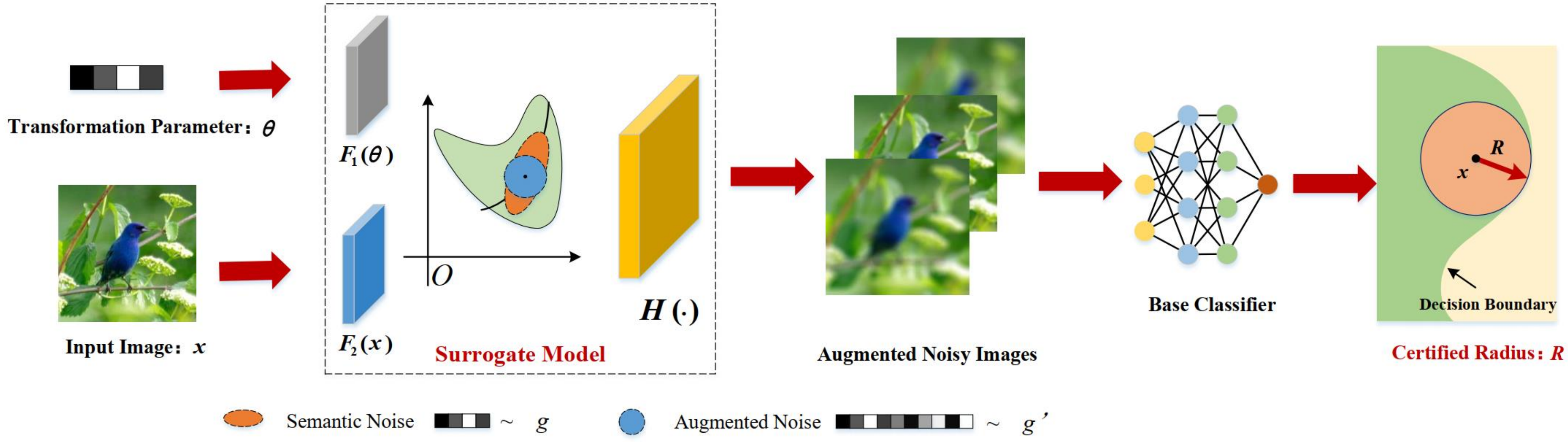
- The generalized smoothed classifier is,

$$\tilde{G}(\tilde{x}) = \mathbb{E}_{\tilde{\theta}\sim\tilde{g}(\cdot)}\left[\tilde{f}(\tilde{\tau}(\tilde{\theta},\tilde{x}))\right], \qquad \tilde{\tau}(\tilde{\theta},\tilde{x}) = \tilde{H}(\tilde{F}_1(\tilde{\theta}) + \tilde{F}_2(\tilde{x})),$$

- We design a surrogate neural network to simulate the transformations

$$\tilde{F}_1(\tilde{\theta}) = \begin{pmatrix} \theta \\ F_1(\theta) + \theta' \end{pmatrix}, \tilde{F}_2(\tilde{x}) = \begin{pmatrix} \mathbf{0}_{d-n} \\ F_2(x) \end{pmatrix}, \quad \tilde{H}(\tilde{x}) = \begin{bmatrix} I_{d-n} & \\ & H(x) \end{bmatrix}.$$

- Illustration of the method



GSmooth: Generalized Randomized Smoothing

- Main theorem for the Generalized Randomized Smoothing

Theorem 2. *Suppose $f(x)$ is a classifier and $\tilde{G}(\tilde{x})$ is the smoothed classifier defined in Eq. (7), if there exist $\underline{p_A}$ and $\overline{p_B}$ satisfying*

$$\tilde{G}(\tilde{x})_A \geqslant \underline{p_A} \geqslant \overline{p_B} \geqslant \tilde{G}(\tilde{x})_B,$$

then $y_A = \arg\max_{i \in \mathcal{Y}} \tilde{G}(\tilde{\tau}(\tilde{\xi}, \tilde{x}))_i$ for any $\|\xi\|_2 \leqslant R$, where

$$R = \frac{1}{2M^*} \int_{\overline{p_B}}^{\underline{p_A}} \frac{1}{\Phi(p)} dp, \tag{10}$$

and the coefficient M^* is defined as

$$M^* = \max_{\xi, \theta \in P} \sqrt{1 + \left\| \frac{\partial F_2(y_\xi)}{\partial \xi} - \frac{\partial F_1(\theta)}{\partial \theta} \right\|_2^2}. \tag{11}$$

Experimental Results

- Certified Accuracy of several types of semantic transformations on CIFAR-10 and CIFAR-100
- Competitive results on resolvable cases
- Non-resolvable cases

Transformation	Type	Dataset	Certified Radius	Certified Accuracy (%) GSmooth (Ours)	TSS	DeepG	Interval	VeriVis	Semanify-NN	IndivSPT/ distSPT
Rotational Blur	Non-resolvable	MNIST	$\|\|\alpha\|\|_2 < 10$	**95.9**	–	–	–	–	–	–
		CIFAR-10	$\|\|\alpha\|\|_2 < 10$	**39.7**	–	–	–	–	–	–
		CIFAR-100	$\|\|\alpha\|\|_2 < 10$	**27.2**	–	–	–	–	–	–
Defocus Blur	Non-resolvable	MNIST	$\|\|\alpha\|\|_2 < 5$	**89.2**	–	–	–	–	–	–
		CIFAR-10	$\|\|\alpha\|\|_2 < 5$	**25.0**	–	–	–	–	–	–
		CIFAR-100	$\|\|\alpha\|\|_2 < 5$	**13.1**	–	–	–	–	–	–
Zoom Blur	Non-resolvable	MNIST	$\|\|\alpha\|\|_2 < 0.5$	**93.9**	–	–	–	–	–	–
		CIFAR-10	$\|\|\alpha\|\|_2 < 0.5$	**44.6**	–	–	–	–	–	–
		CIFAR-100	$\|\|\alpha\|\|_2 < 0.5$	**14.2**	–	–	–	–	–	–
Pixelate	Non-resolvable	MNIST	$\|\|\alpha\|\|_2 < 0.5$	**87.1**	–	–	–	–	–	–
		CIFAR-10	$\|\|\alpha\|\|_2 < 0.5$	**45.3**	–	–	–	–	–	–
		CIFAR-100	$\|\|\alpha\|\|_2 < 0.5$	**30.2**	–	–	–	–	–	–

Experimental Results

- Empirical accuracy under adaptive attacks

Type	Certified Acc. (%)	Adaptive Attack Acc. (%)	
	GSmooth		Vanilla
Gaussian blur	67.4	68.1	3.4
Translation	82.2	87.5	4.2
Brightness	82.5	85.9	9.6
Rotation	65.6	68.4	65.4
Rotational blur	39.7	45.0	33.1
Defocus blur	25.0	25.5	16.6
Pixelate	45.3	49.2	38.2

- Empirical accuracy on subsets of CIFAR-10-C

Type	AugMix	TSS	GSmooth
Gaussian blur	67.4	75.8	**76.0**
Brightness	**82.4**	71.8	72.1
Defocus blur	72.2	75.6	**76.8**
Zoom blur	70.8	75.2	**77.1**
Motion blur	68.6	70.2	**70.5**
Pixelate	50.9	76.0	**76.7**

Thanks !