Randomized Smoothing Under Attack: how good is it in practice?

Thibault MAHO, Teddy FURON, Erwan LE MERRER
Univ. Rennes, Inria, CNRS, IRISA, Rennes, France

4159

Goal

Consider Randomized Smoothing as a defense and to evaluate its effectiveness on black-box attacks

Randomized Smoothing

- For a binary classifier f, creates a deterministic classifier:

$$g_\sigma(\mathbf{x}) = \arg\max_{y\in\{0,1\}} \mathbb{P}[f(\mathbf{x}+\sigma\mathbf{N}) = y],\ \mathbf{N}\sim\mathcal{N}(0,I).$$

- g_σ have a certified local robustness

$$R(\mathbf{x},\sigma) = \sigma\Phi^{-1}(P[f(\mathbf{x}+\sigma\mathbf{N}) = g_\sigma(\mathbf{x})], \mathbf{N}\sim\mathcal{N}(0,I)$$

→ All points at a distance from *x* lower than $R(\mathbf{x},\sigma)$ are classified with the same label

- In practice, uses Monte Carlo with *n* samples to estimate R
→ in practice equivalent to the random classifier $g_{\sigma,n}$

Problem

- $g_{\sigma,n}$ **is not deterministic**
- Recommendation not clear:
  - Number of samples: the higher the better, but no consensus for the minimum
  - Amount of noise: A high σ gives a better certification, but leads to an accuracy drop.
- Attacks are not considered in the literature anymore

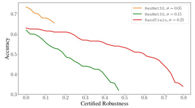


Fig. 1: Certification for ResNet50

Adversarial Examples with Random Classifier

- Classical Definition defined on a deterministic classifier
- On random classifier, they have a confidence score P_a.

$$\mathbb{P}[g_{\sigma,n}(\mathbf{x}_a) \neq g_\sigma(\mathbf{x}_o)] \geq P_a.$$

Impact on black-box attacks

- Black-box attacks main steps:
  - Binary Search
  - Gradient Estimation
- Randomized Smoothing have no impact on gradient estimation but can perturb the binary search where a single wrong prediction lead to a bad convergence.
→ The lower the number of samples, the more the binary search can be impacted

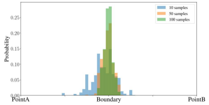
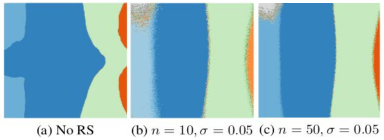


Fig. 3: Distribution of the output of a binary search with RS.

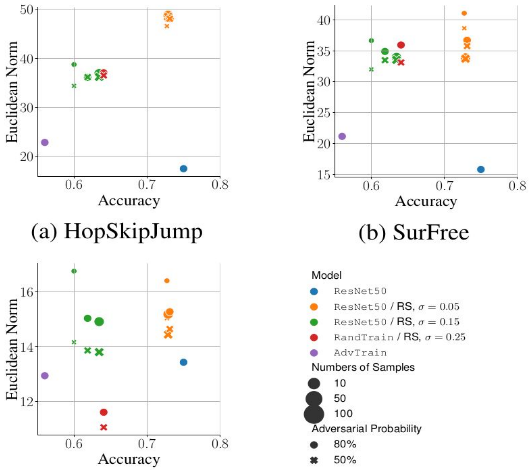
- The reason: Randomized Smoothing greatly perturbs the boundary with a low number of samples.
The prediction of a point on the boundary easily changes.



(a) No RS (b) $n = 10, \sigma = 0.05$ (c) $n = 50, \sigma = 0.05$

Fig. 2: 2D slice in the image space of ResNet50 with and without RS. Each point is an image, his color represents the elected label

Results



(a) HopSkipJump (b) SurFree

(c) RayS

- The results of all black-box attacks are impacted
- Practical distortions at least 30 times larger than the certified robustness
- Comparison between the recommendations made for theoretical robustness and practical robustness:

Theoretical Robustness	Practical Robustness
• Many Queries • High amount of noise	• Few number of queries is enough • Small amount of noise is enough

J. Chen et al, “Rays: A ray searching method for hard-label adversarial attack”. In SIGKBB 2020; T. Maho et al, “SurFree: a fast surrogate-free black-box attack”. In CVPR 2021; J. Chen et al, “HopSkipJumpAttack: A Query-Efficient Decision-Based Attack”