

ANITI

Communauté
d'universités
et établissements
de Toulouse



Fast Certificates for Semantic Segmentation

CALM Chair

Thomas MASSENA, supervised by Pr. Mathieu SERRURIER and Dr. Corentin FRIEDRICH.

Deep Learning models can be adversarially perturbed such that they exhibit *unsafe* behaviour.

Such a phenomenon is a problem, preventing the use of Deep Learning-based models in safety-critical scenarios.

Most methods are tailored and tested only on **classification tasks**.

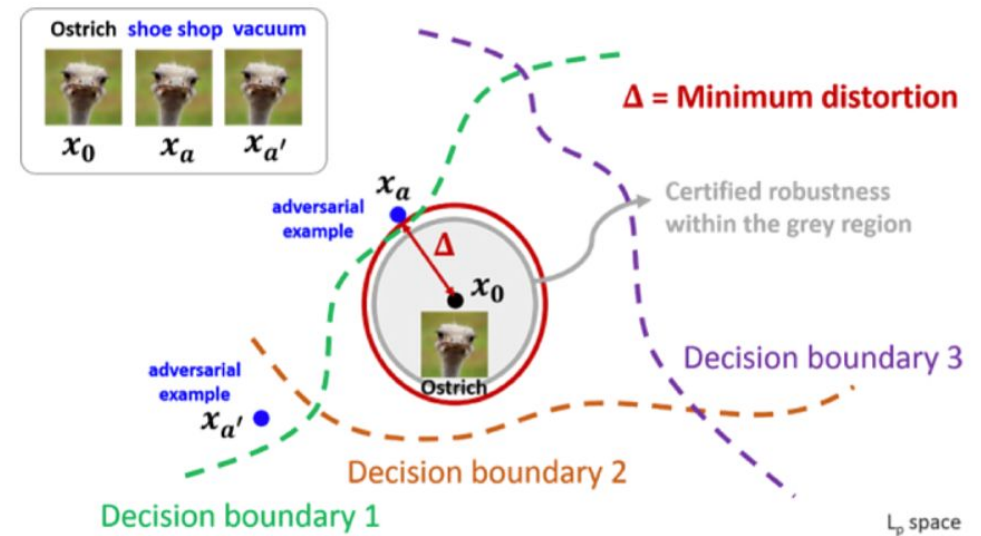


Figure courtesy of : *Getting CLEVER(er): Expanding the Scope of a Robustness Metric for Neural Networks* (Medium)

We can distinguish two different types of desirable robustness certificates. We provide general formulations.

- Agnostic from performance measures.
- Allows for arbitrary degradation objectives.

We use 1-Lipschitz constrained networks for efficiency at testing time !

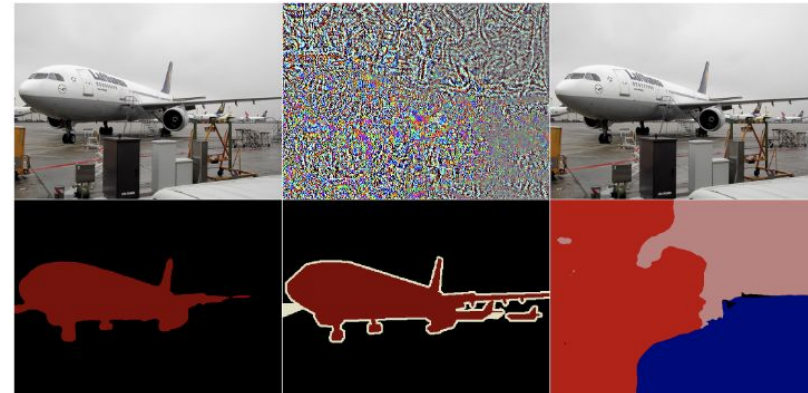
$$f : \mathbb{R}^n \rightarrow \mathbb{R}^K, \forall x, y \in \mathbb{R}^n, \|f(x) - f(y)\| \leq \|x - y\|$$



I want to drop the pixel accuracy by 30 % !



I can apply a noise of norm 0.25, how bad can the pixel accuracy get ?



Only randomized smoothing approaches currently exist to ensure the robustness of pixel classifications for segmentation on big neural networks.

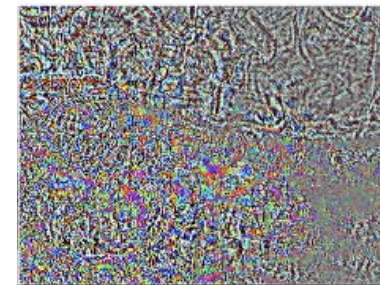
These methods give probabilistic robustness certificates for pixel-wise classifications up to a robustness radius R .

The main problem with these methods is their limited efficiency. Making real-life deployment impossible.

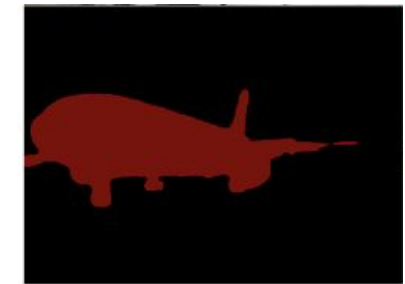
Fischer et al., "Scalable Certified Segmentation via Randomized Smoothing", ICML 2021.



+



Run N times



Compute $H \times W$
p-values

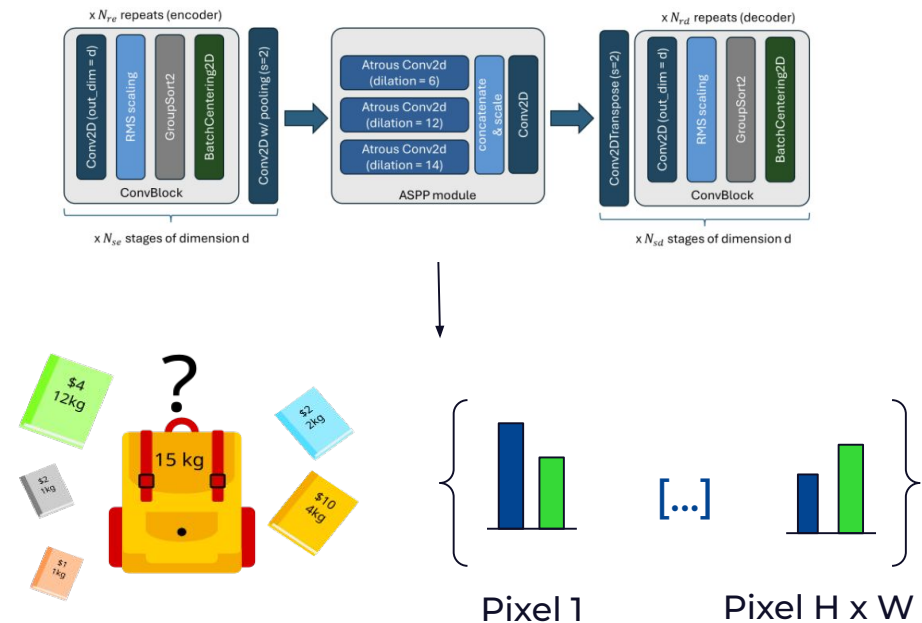
+

Apply Hölm correction

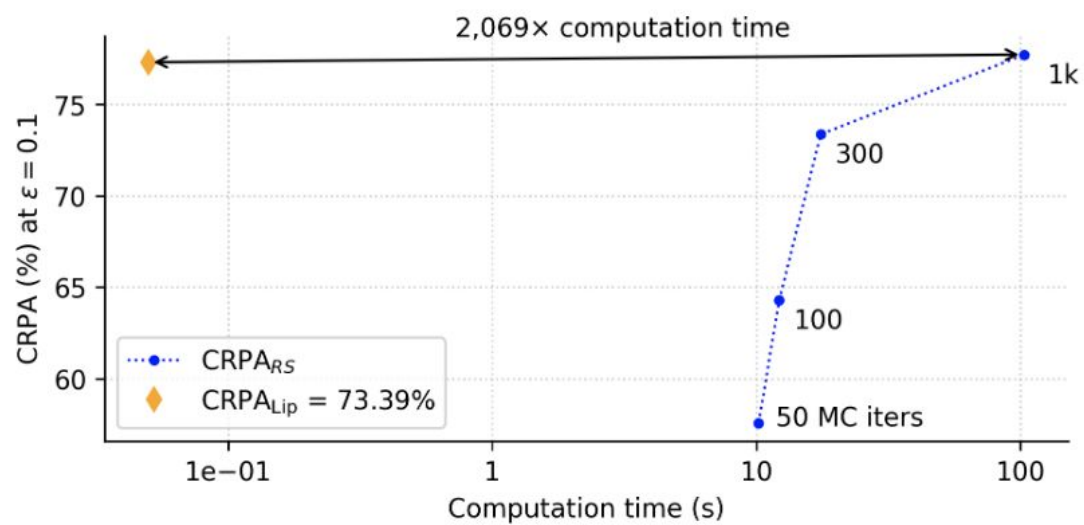
If the neural network is K-Lipschitz wrt the output vector. The certification of pixels can be related to a Knapsack problem. Where:

- Each pixel has a cost.
- Each pixel has a profit.
- The attacker has a bounded budget.

We devise efficient strategies to enable certification of different metrics (e.g. the FNR, IoU, etc...).



On the same neural network:



Compared on the Oxford IIIT Pets dataset.

On different neural networks:

ϵ	Method	CRPA	Time (total / nb samples)	# forward passes / sample
0.1	Lipschitz bound (ours)	81.80%	≈ 0.1 s	1
0.1	SEGCERTIFY ($\sigma = 0.3$)	$53.48 \pm 0.59\%$	59.8 s $\times 594$	60
0.1	SEGCERTIFY ($\sigma = 0.2$)	$83.13 \pm 0.33\%$	62.1 s $\times 624$	80
0.17	Lipschitz bound (ours)	77.34%	≈ 0.1 s	1
0.17	SEGCERTIFY ($\sigma = 0.4$)	$38.91 \pm 0.53\%$	60.3 s $\times 594$	60
0.17	SEGCERTIFY ($\sigma = 0.2$)	$84.84 \pm 0.73\%$	63.3 s $\times 683$	120

Compared on the CityScapes dataset.