Introduction to Machine and Deep Learning Theory

Certified Robustness II: Ablations on Randomized Smoothing. High Dimensions and Computer Vision

Aleksandr Petiushko

Lomonosov MSU
Faculty of Mechanics and Mathematics

April 26th, 2023

AP

Content

1. Certified robustness: recap
2. Certification in High Dimensional case
3. Certification of Semantic Perturbations

AP

Certified Robustness: for classification

- Let us NN function $f(x)$ is the classifier to K classes: $f : \mathbb{R}^d \to Y, Y = \{1, \dots, K\}$
- Usually we have NN $h(x) : \mathbb{R}^d \to \mathbb{R}^K$, and $f(x) = \arg\max_{i \in Y} h(x)_i$

Deterministic approach

Need to find the class of input perturbation $S(x, f)$ so as the classifier's output doesn't not change, or more formally:

$$f(x + \delta) = f(x) \quad \forall \delta \in S(x, f)$$

Probabilistic approach

Need to find the class of input perturbation $S(x, f, P)$ w.r.t. robustness probability P s.t.:

$$Prob_{\delta \in S(x,f,P)}(f(x + \delta) = f(x)) = P$$

Remark: Probabilistic approach coincides with Deterministic one when $P = 1$.

AP

Certified Robustness: for regression

- Let us NN function $f(x)$ NN f(x) is the regressor: $f : \mathbb{R}^d \to \mathbb{R}$

Deterministic approach

Need to find the class of input perturbation $S(x, f, f_{low}, f_{up})$ w.r.t. the upper and lower bounds on the output perturbation f_{low}, f_{up} s.t.:

$$f(x) - f_{low} \leq f(x+\delta) \leq f(x) + f_{up} \quad \forall \delta \in S(x, f, f_{low}, f_{up})$$

Probabilistic approach

Need to find the class of input perturbation $S(x, f, f_{low}, f_{up}, P)$ w.r.t. robustness probability P and the upper / lower bounds on the output perturbation f_{low}, f_{up} s.t.:

$$Prob_{\delta \in S(x, f, f_{low}, f_{up}, P)}(f(x) - f_{low} \leq f(x+\delta) \leq f(x) + f_{up}) = P$$

AP

Certified Robustness: inverse tasks for classification

- Suppose that we know the input perturbation class S
- For classification we have only probabilistic formulation

Classification

Need to measure the probability P of retaining the classifier's output under some class of input perturbations S:

$$Prob_{\delta \in S}(f(x+\delta) = f(x)) = P$$

Certified Robustness: inverse tasks for regression

- Suppose that we know the input perturbation class S
- For regression we have both deterministic and probabilistic formulations

Regression (deterministic formulation)

Need to find the upper and lower bounds $f_{low}(f, x, S)$, $f_{up}(f, x, S)$ of the output perturbation under some class of input perturbations S:

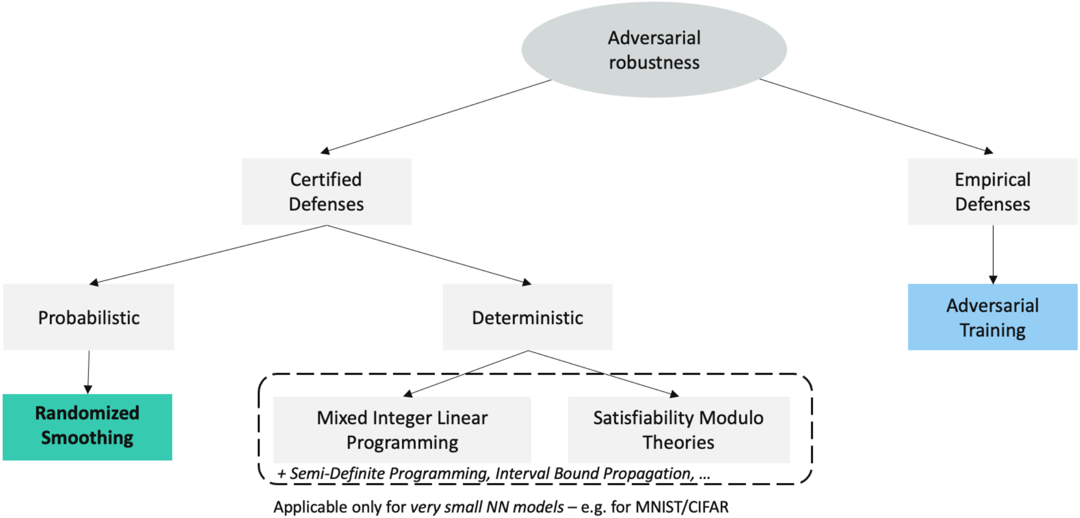
$$f(x) - f_{low}(f, x, S) \leq f(x + \delta) \leq f(x) + f_{up}(f, x, S)$$

Regression (probabilistic formulation)

Need to measure the probability P of keeping the classifier's output inside the lower / upper bounds f_{low}, f_{up} under some class of input perturbations S:

$$Prob_{\delta \in S}(f(x) - f_{low} \leq f(x + \delta) \leq f(x) + f_{up}) = P$$

Adversarial Robustness: overview



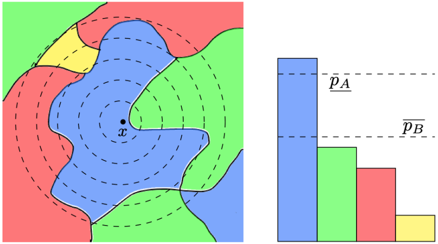
AP

Randomized Smoothing

Idea of Randomized Smoothing (RS)

- Let's use the **T**est **T**ime **A**ugmentation (**TTA**) in order to mitigate the boundary effect
- The new classifier $g(x)$ is defined as:

$$g(x) = \arg\max_{c \in Y} P(f(x+\epsilon) = c), \epsilon \sim N(0, \sigma^2)$$



RS main result

- If the initial classifier $f(x)$ is robust under Gaussian noise,
- Then the new classifier $g(x)$ is robust under **ANY** noise

AP

Randomized Smoothing: Theory overview

Theorem: Certification Radius

Suppose $c_A \in Y$ and $\underline{p_A}, \overline{p_B} \in [0,1]$ satisfy
$\mathbb{P}(f(x+\epsilon) = c_A) \geq \underline{p_A} \geq \overline{p_B} \geq \max_{c_B \neq c_A} \mathbb{P}(f(x+\epsilon) = c_B)$. Then
$g(x+\delta) = c_A \quad \forall \|\delta\|_2 < R$, where

$$R = \frac{\sigma}{2}(\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B}))$$

Remark. Φ^{-1} is the inverse of the standard Gaussian CDF: $\Phi(x) = \frac{1}{2\pi}\int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$.

Randomized Smoothing: Better training

- For an original RS the training process is just augmentation with Gaussian Noise images
- Idea of **SmoothAdv**[1]: let's do adversarial training (AT) using attacks on smoothed classifier $g(x)$!
  - Original RS hard example (in the vicinity ϵ):

$$x' = \arg\max_{x':\|x'-x\|_2 \le \epsilon} \mathbb{E}_{\delta \sim N(0,\sigma^2 I)}[L(f_\theta(x'+\delta), y)]$$

  - SmoothAdv hard example:

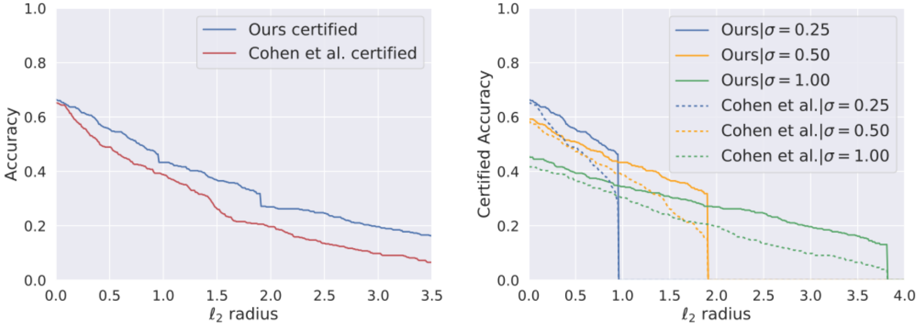
$$x' = \arg\max_{x':\|x'-x\|_2 \le \epsilon} L(g_\theta(x'), y) = \arg\max_{x':\|x'-x\|_2 \le \epsilon} L(\mathbb{E}_{\delta \sim N(0,\sigma^2 I)}[f_\theta(x'+\delta)], y)$$

  - If cross entropy loss is used, then the difference is that we changing $\sum \log$ to $\log \sum \Rightarrow$ using Jensen's inequality, we have $\sum \log \le \log \sum$

[1] Salman, Hadi, et al. "Provably robust deep learning via adversarially trained smoothed classifiers." 2019

AP

Randomized Smoothing: SmoothAdv Results



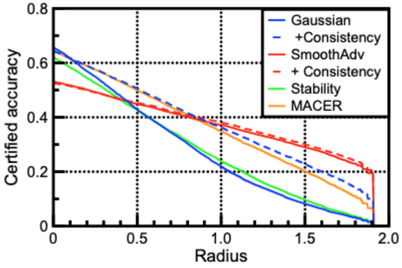
ℓ_2 RADIUS (IMAGENET)	0.5	1.0	1.5	2.0	2.5	3.0	3.5
COHEN ET AL. [6] (%)	49	37	29	19	15	12	9
OURS (%)	**56**	**45**	**38**	**28**	**26**	**20**	**17**

AP

Randomized Smoothing: Regularization

- What if we can work on top of smoothed classifier $g(x)$ to make it more reasonable?
- Idea of Consistency Regularization[2]: let's force *similarity* between *smoothed* and *perturbed* predictions as well as minimizing the entropy of smoothed output:

$$L_{CR}(x) = \lambda \mathbb{E}_{\delta \sim N(0,\sigma^2 I)} D_{KL}(g(x)||f(x+\delta)) + \eta H(g(x))$$



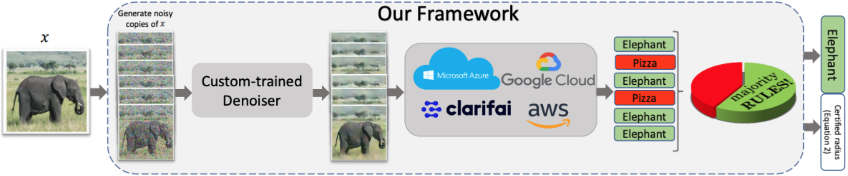
(b) $\sigma = 0.50$

[2] Jeong, Jongheon, and Jinwoo Shin. "Consistency regularization for certified robustness of smoothed classifiers." 2020

Randomized Smoothing: Black-box access

- What if we **cannot change the pretrained classifier**, but want to increase its certified robustness?
- Idea of **Black-box smoothing**[3]: Let's train a **denoiser** D used after we've added Gaussian noise!
  - And then simply apply the majority rule

$$g(x) = \arg\max_{c \in Y} \mathbb{P}[f(D(x+\delta)) = c], \quad \delta \sim N(0, \sigma^2 I)$$



[3] Salman, Hadi, et al. "Black-box smoothing: A provable defense for pretrained classifiers." 2020

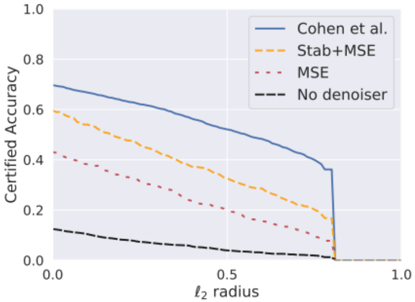
Randomized Smoothing: Denoiser for Black-box

- Denoiser: trained with two losses for every Gaussian σ:
  - MSE
  - Stability (classification cross entropy)

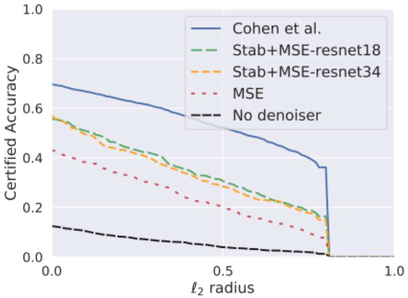
(a) Noisy (b) MSE (c) Stab+MSE

Randomized Smoothing: Results for Black-box

Full access to classifier during training



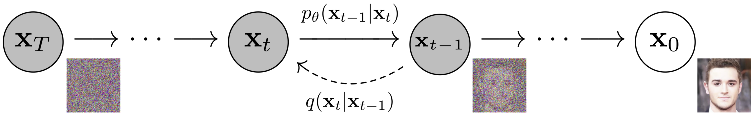
Using surrogate classifier during training



ℓ_2 Radius (ImageNet)	0.25	0.5	0.75	1.0	1.25	1.5
Cohen et al. (2019) (%)	(70)62	(70)52	(62)45	(62)39	(62)34	(50)29
No denoiser (baseline) (%)	(49)32	(12)4	(12)2	(0)0	(0)0	(0)0
Ours (Black-box) (%)	(69)48	(56)31	(56)19	(34)12	(34)7	(30)4
Ours (White-box) (%)	(67)50	(60)33	(60)20	(38)14	(38)11	(38)6

Denoiser by DDPM[5]

- A novel approach[4] to use off-the-shelf models:
  - SotA classifier (trained on clean images)
  - Denoising Diffusion Model
    - Based on the noise level σ, estimate $\bar{\alpha}_t, t$
    - Generate $x_t \sim N(\sqrt{\bar{\alpha}_t} \cdot x, (1 - \bar{\alpha}_t) I)$
    - Denoise by DDPM decoder (using **only 1 step**): $\hat{x} = denoise(x_t)$
    - Classify!
- Results in 14% improvement over the prior certified SoTA, and an improvement of 30% over denoised smoothing



[4] N. Carlini, F. Tramer, and Z. Kolter. "(Certified!!) Adversarial Robustness for Free!", 2022

[5] J. Ho, A. Jain, and P. Abbeel. "Denoising diffusion probabilistic models", 2020

Randomized Smoothing: vector functions

- Previously all results were for the classifiers: $f, g : \mathbb{R}^d \to Y, Y = \{1, \ldots, K\}$, $g(x) = \arg\max_{c \in Y} P(f(x+\epsilon) = c), \epsilon \sim N(0, \sigma^2)$
- Let's consider the vector-based functions f (e.g., feature vector): $\mathbb{R}^d \to \mathbb{R}^D$
- Then the smoothed version g of it we'll define as: $g(x) = \mathbb{E}_{\epsilon \sim N(0, \sigma^2 I)}[f(x+\epsilon)]$
- In this case the following relation to Lipschitz functions can be established[6]:

Lipschitz-continuity of smoothed vector function

Suppose that $g(x)$ is continuously differentiable for all x. If for all x, $\|f(x)\|_2 = 1$, then $g(x)$ is $L-$Lipschitz in l_2-norm with $L = \sqrt{\frac{2}{\pi\sigma^2}}$.

[6] Pautov, Mikhail, et al. "Smoothed Embeddings for Certified Few-Shot Learning." 2022

Randomized Smoothing: adversarial embedding risk

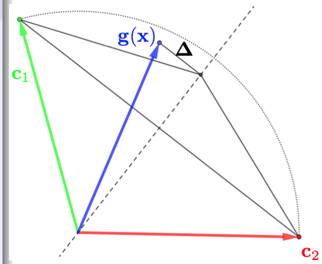
- Let's establish the beautiful geometrical fact useful for the few-shot classification:

Adversarial embedding risk

Given an input $x \in \mathbb{R}^d$ and the embedding $g : \mathbb{R}^d \to \mathbb{R}^D$ the closest point on to decision boundary in the embedding space is located at a distance:

$$\gamma = \|\Delta\|_2 = \frac{\|c_2 - g(x)\|_2^2 - \|c_1 - g(x)\|_2^2}{2\,\|c_2 - c_1\|_2^2},$$

where $c_1 \in \mathbb{R}^D$ and $c_2 \in \mathbb{R}^D$ are the two closest prototypes.



- γ is the distance between classifying embedding and the decision boundary between classes represented by c_1 and c_2.
- γ is the minimum l_2-distortion in the embedding space required to change the prediction of g.

Randomized Smoothing: certification

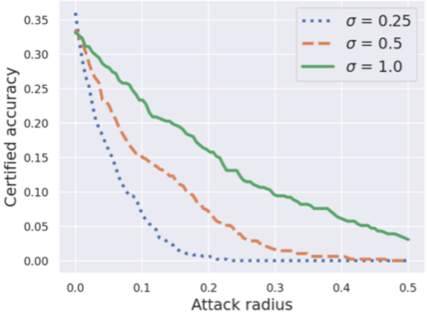
- Two results above lead to the certification guarantee:

Robustness guarantee

Certified radius r of g at x, where g is the smoothed version of $f : \|f(x)\|_2 = 1$, is

$$r = \frac{\gamma}{L}$$

1-shot results for *mini*ImageNet[7]



[7] Vinyals, Oriol, et al. "Matching networks for one shot learning." 2016

Randomized Smoothing: Some Results on Regression

- Let function $f : \mathbb{R}^d \to \mathbb{R}$
- Smoothed version $g : \mathbb{R}^d \to \mathbb{R}$, $g(x) = \mathbb{E}_{\epsilon \sim N(0,\sigma^2)}[f(x+\epsilon)]$
- Φ^{-1} is the inverse of the standard Gaussian CDF: $\Phi(x) = \frac{1}{2\pi} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$
- Result: the solution[8] for inverse certification problem (deterministic formulation):

Theorem

For any bounded $f : \mathbb{R}^d \to [l, u]$, the map $\eta(x) = \sigma\Phi^{-1}(\frac{g(x)-l}{u-l})$ is 1-Lipschitz, implying

$$l + (u-l) \cdot \Phi(\frac{\eta(x) - \|\delta\|_2}{\sigma}) \leq g(x+\delta) \leq l + (u-l) \cdot \Phi(\frac{\eta(x) + \|\delta\|_2}{\sigma})$$

Exercise. Prove it.

[8] Chiang, Ping-yeh, et al. "Detection as regression: Certified object detection with median smoothing." 2020

Randomized Smoothing: norms

- Randomized Smoothing = Smoothing distribution + **norm** l_p of perturbation
- Using l_p-balls is neither necessary nor sufficient for perceptual robustness
- Certification is only for much smaller regions than humans can do
- Remark about physical nature of l_p-balls:
  - l_2 corresponds to the power of signals
  - l_1 corresponds to the pixel mass
  - l_∞ corresponds to the noise in camera sensors
  - l_0 corresponds to the practical patch robustness

AP

Randomized Smoothing: High Dimensional Case

- The perturbation δ is measured by l_p-norm
- $p = 1$ and $p = 2$ are the only **special cases**[9]: $R = \frac{\sigma}{2}(\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B}))$
- Unfortunately, these are **only** examples of **non-decreasing** with **input dimension** d
- For any $p \geq 2$, the certification radius[10] is decreasing with dimensionality d:

$$R_p(x) = \frac{\sigma}{2d^{\frac{1}{2}-\frac{1}{p}}}(\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B}))$$

- And the most important case for Computer Vision (CV), $p = \infty$, means

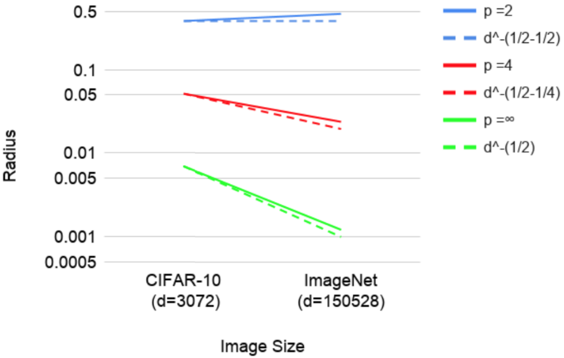
$$R_\infty \sim \frac{1}{\sqrt{d}}$$

[9] Yang, Greg, et al. "Randomized smoothing of all shapes and sizes." 2020

[10] Kumar, Aounon, et al. "Curse of dimensionality on randomized smoothing for certifiable robustness." 2020

AP

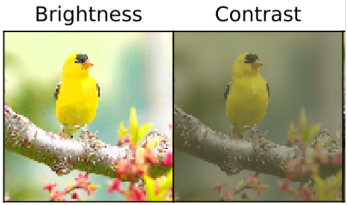
Randomized Smoothing: CV illustration



AP

High Dimension Case in CV

- Any **semantic-meaningful** perturbation in **CV** leads to **high l_∞-perturbation**, and the dimension of an image $d = H \times W$ usually is very high (like millions of pixels)
- $R_\infty \sim \frac{1}{\sqrt{d}}$ means that there is **no any practical certified radius**
- E.g., for semantic-specific transformations like **contrast** and **brightness** the **error is higher** than on clean images up to 50-60% on *Common Corruptions*[11] on ImageNet



Network	Error	Bright	Contrast
AlexNet	43.5	100	100
SqueezeNet	41.8	97	98
VGG-11	31.0	75	86
VGG-19	27.6	68	80
VGG-19+BN	25.8	61	74
ResNet-18	30.2	69	78
ResNet-50	23.9	57	71

[11] Hendrycks, Dan, and Thomas Dietterich. "Benchmarking neural network robustness to common corruptions and perturbations." 2019

High Dimension Case in CV: Autonomous Driving

- The same is true for safety-critical applications like autonomous driving[12]

contrast(1.8) original brightness(50)

Transformation	#err
Brightness	97
Contrast	31

[12]Tian, Yuchi, et al. "Deeptest: Automated testing of deep-neural-network-driven autonomous cars." 2017

Semantic perturbations for additive parameters

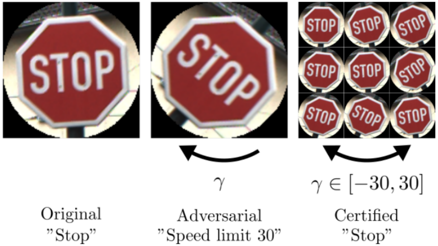
- So... let's certify semantic perturbations[13]!
  - Usually parameterized by a much smaller dimension (1 or 2 dimensional)
- Consider **rotations** and **translations** γ_β parameterized by β: $\gamma_\beta : \mathbb{R}^d \to \mathbb{R}^d$
- A smoothed classifier $g(x) = \arg\max_{c \in Y} P_{\beta \sim N(0,\sigma^2)}(f \circ \gamma_\beta(x) = c)$
- Also **interpolation** procedure is taken into account because after rotation we need to interpolate anyway

Certification Radius

Suppose $c_A \in Y$ and $\underline{p_A}, \overline{p_B} \in [0,1]$ satisfy $\mathbb{P}_{\beta \sim N(0,\sigma^2)}(f \circ \gamma_\beta(x) = c_A) \geq \underline{p_A} \geq \overline{p_B} \geq \max_{c_B \neq c_A} \mathbb{P}_{\beta \sim N(0,\sigma^2)}(f \circ \gamma_\beta(x) = c_B)$. Then $g \circ \gamma_\beta(x) = c_A \quad \forall \|\gamma\|_2 < r_\gamma$, where $r_\gamma = \frac{\sigma}{2}(\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B}))$.

[13] Fischer, Marc, et al. "Certified defense to image transformations via randomized smoothing." 2020

Semantic perturbations for additive parameters: results

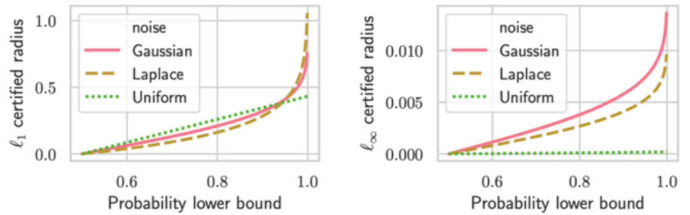


Rotation						r_γ percentile		
Dataset	$\mathcal{I}$	σ_γ	α_γ	f Acc.	g Acc.	25^{th}	50^{th}	75^{th}
ImageNet	bil.	10	0.001	0.39	0.29	10.81	10.81	10.81
ImageNet	bil.	10	0.001	0.39	0.29	18.29	18.29	18.29
ImageNet	bil.	30	0.001	0.39	0.28	9.09	16.59	28.60
ImageNet	bil.	30	0.001	0.39	0.28	20.22	25.36	$30^{\dagger}$
ImageNet	bic.	10	0.001	0.39	0.29	10.40	10.40	10.40
ImageNet	bic.	30	0.001	0.39	0.27	9.33	17.00	28.74
ImageNet	near.	10	0.001	0.39	0.29	9.62	9.62	9.62
ImageNet	near.	30	0.001	0.39	0.26	7.38	16.63	27.72

Translation						r_γ percentile		
Dataset	$\mathcal{I}$	σ_γ	α_γ	f Acc.	g Acc.	25^{th}	50^{th}	75^{th}
ImageNet	bil.	50	0.001	0.48	0.36	2.4%	2.4%	2.4%
ImageNet	bic.	50	0.001	0.48	0.36	2.4%	2.4%	2.4%

Randomized Smoothing: smoothing distribution

- Randomized Smoothing = Smoothing **distribution** + norm l_p of perturbation
- Original (and most of the follow-up ones) work uses Gaussian Smoothing
- Other types of randomized smoothing could be taking into account: e.g. Uniform[14] or Laplacian[15]
- What about other types?



[14] Lee, Guang-He, et al. "Tight certificates of adversarial robustness for randomly smoothed classifiers." 2019

[15] Teng, Jiaye, et al. "ℓ_1 Adversarial Robustness Certificates: a Randomized Smoothing Approach." 2019

Semantic perturbations and multiplicative parameters

- All research above is concentrated on **additive** perturbations
- Let's investigate the **multiplicative** parameters[16] (e.g., *gamma correction* $G_\gamma(x) = x^\gamma$ in CV)
- **Definition**: A parameterized map $\psi_\delta : X \to X,\ \delta \in \mathcal{B} \subset \mathbb{R}^n$ is called multiplicatively composable if $(\psi_\delta \circ \psi_\theta)(x) = \psi_{(\delta \cdot \theta)}(x),\ \forall x \in X,\ \forall \delta, \theta \in \mathcal{B}$
- Example: $G_\beta \circ G_\gamma(x) = (x^\gamma)^\beta = x^{\gamma \cdot \beta} = G_{\gamma \cdot \beta}(x)$

(a) $\gamma = 0.5$ (b) $\gamma = 1$ (c) $\gamma = 2$

[16] Muravev, Nikita, and Aleksandr Petiushko. "Certified Robustness via Randomized Smoothing over Multiplicative Parameters." 2021

Semantic perturbations and multiplicative parameters: results

- To work under this limitation, the new type of smoothing distribution is needed:
  - Positive support
  - Mean at 1
- The proposal to use is **Rayleigh** distribution: $p_\beta(z) = \sigma^{-2} z e^{-z^2/(2\sigma^2)}, z \geq 0$
- Then the following is true: $g \circ \psi_\gamma(x) = c_A$ for all γ satisfying $\gamma_1 < \gamma < \gamma_2$, where γ_1, γ_2 are the only solutions of the following equations:
  $F(\gamma_1^{-1} F^{-1}(\overline{p_B})) + F(\gamma_1^{-1} F^{-1}(1 - \underline{p_A})) = 1,$
  $F(\gamma_2^{-1} F^{-1}(\underline{p_A})) + F(\gamma_2^{-1} F^{-1}(1 - \overline{p_B})) = 1,$
  and $F(z) = 1 - e^{-z^2/(2\sigma^2)}$ is the CDF of γ.
- The results are better for $\gamma < 1$ in comparison to Uniform, Gaussian and Laplace smoothing

$\underline{p_A}$	$\overline{p_B}$	γ_1	γ_2
0.600	0.400	0.86	1.15
	0.200	0.71	1.33
0.700	0.300	0.72	1.32
	0.100	0.54	1.56
0.800	0.200	0.57	1.52
0.900	0.100	0.39	1.82
0.990	0.010	0.12	2.58
0.999	0.001	0.04	3.16

AP

Semantic perturbations and compositions

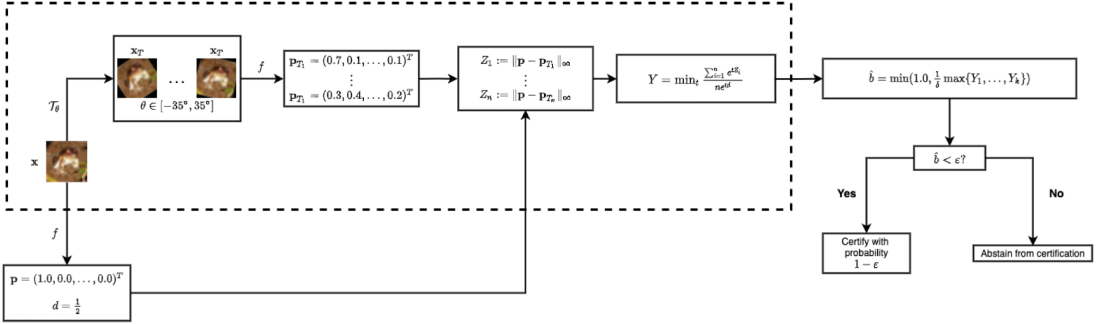
- Usually multiple transformations are applied to the input: how to certify the composition?
- Forward theoretical estimation is difficult $\Rightarrow$ let's try inverse (probabilistic) task[17]!
- The proposal to use Chernoff-Cramer inequality[18] (Markov's inequality corollary) to provide the statistically-grounded **estimations for the certification**, where perturbed **radius** is **already given**
- Can be easily used for **any** semantic **perturbation** and **any compositions**

Dataset	Transform	Parameters	Training type	ERA	PCA(ε)		
					$\varepsilon = 10^{-10}$	$\varepsilon = 10^{-7}$	$\varepsilon = 10^{-4}$
	Brightness	$\theta_b \in [-40\%, 40\%]$	plain	58.4%	47.8%	51.6%	55.2%
			smoothing	65.0%	55.4%	59.4%	61.8%
	Contrast	$\theta_c \in [-40\%, 40\%]$	plain	91.6%	62.4%	67.0%	69.6%
			smoothing	88.0%	67.0%	72.8%	74.2%
	Rotation	$\theta_r \in [-10^\circ, 10^\circ]$	plain	73.4%	64.6%	69.0%	71.0%
			smoothing	72.4%	57.4%	63.6%	67.4%
	Contrast + Brightness	see Contrast & Brightness	plain	0.0%	0.0%	0.0%	0.0%
			smoothing	0.4%	0.0%	0.0%	0.0%
	Rotation + Brightness	see Rotation & Brightness	plain	22.6%	16.2%	20.6%	21.8%
			smoothing	30.4%	21.2%	24.6%	27.6%
	Scale + Brightness	see Scale & Brightness	plain	10.2%	10.4%	10.4%	10.4%
			smoothing	41.8%	40.6%	40.6%	40.6%

[17] Pautov, Mikhail, et al. "CC-Cert: A probabilistic approach to certify general robustness of neural networks." 2021

[18] Boucheron, Stéphane, et al. "Concentration inequalities." 2003

Inverse certification for any transformation[19]

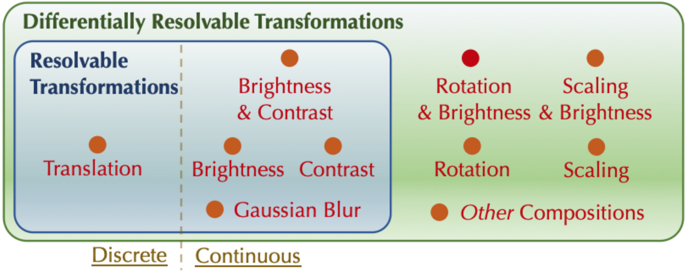


[19]Pautov, Mikhail, et al. "CC-Cert: A probabilistic approach to certify general robustness of neural networks." 2021

AP

Semantic perturbations: further development (1)

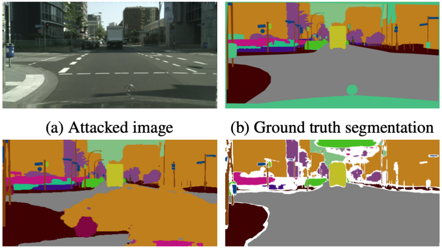
- Later works introduced approaches to take into account different types of perturbations and interpolation errors[20]



[20] Li, Linyi, et al. "Tss: Transformation-specific smoothing for robustness certification." 2020

Semantic perturbations: further development (2)

- Later works introduced approaches to apply certified robustness for other types of CV tasks — e.g. detection[21] and segmentation[22].



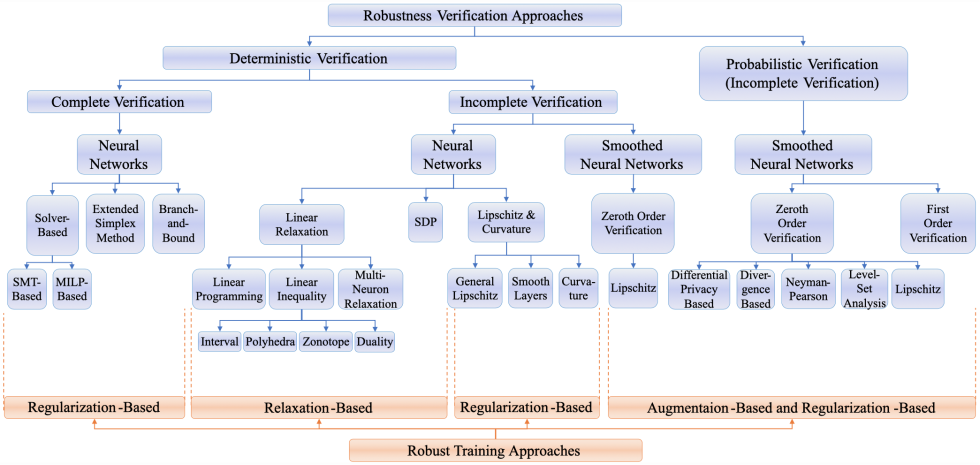
(a) Attacked image (b) Ground truth segmentation

(c) Attacked segmentation (d) Certified segmentation

[21] Chiang, Ping-yeh, et al. "Detection as regression: Certified object detection with median smoothing." 2020

[22] Fischer, Marc, et al. "Scalable certified segmentation via randomized smoothing." 2021

Systematization of Knowledge[23]



[23] SoK, Benchmark and Leaderboard

Takeaway notes

- Straightforward certification in l_∞ is not working for high dimension input
- In Computer Vision no need in any l_p (aside from l_0 for patch attacks, but it is usually also combined with other perturbations)
- Semantic perturbations are much harder to certify (+ interpolation!)
- **Current challenge**: 3D and even non-rigid transformations of **real world**

AP

Thank you!

AP