Fraunhofer IKS

Fraunhofer-Institut für Kognitive Systeme IKS

AISafety-SafeRL 2023 Workshop (IJCAI)
August 19–21, 2023, Macao, SAR, China

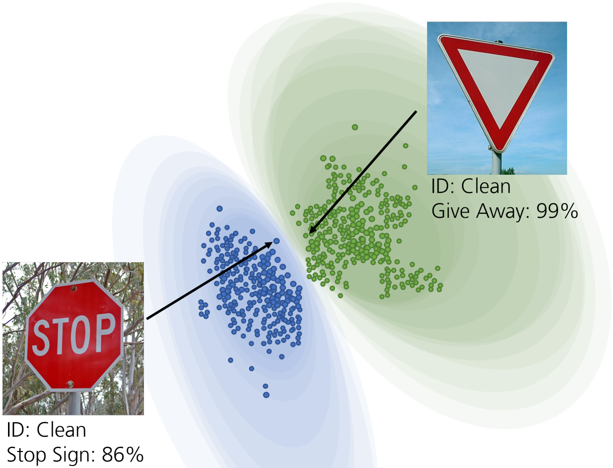
Diffusion Denoised Smoothing for Certified and Adversarial Robust Out-Of-Distribution Detection

Nicola Franco, Daniel Korth, Jeanette Miriam Lorenz, Karsten Rocher, Stephan Günnemann

Background - Adversarial Attacks for In- and Out-of-Distribution Samples

Notation

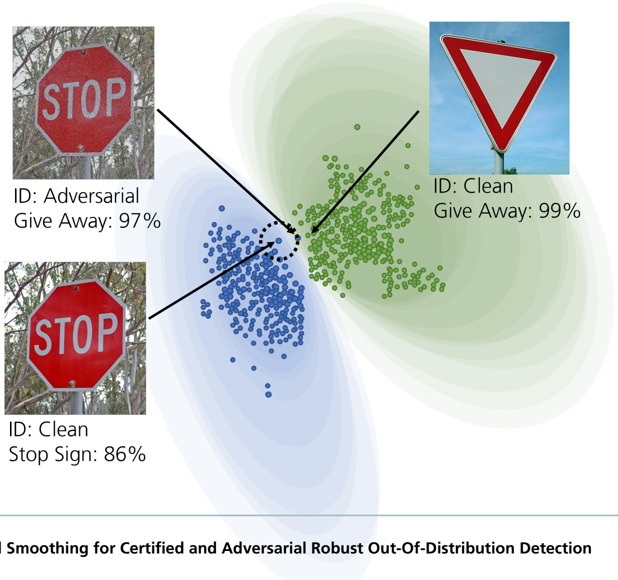
- Classifier $f(x)$
- Confidence, e.g. $h(x) = \mathrm{SoftMax}(f(x))$
- In-distribution (ID) $x \sim \mathcal{D}_{in}$ ($\bullet$, $\bullet$)



Background - Adversarial Attacks for In- and Out-of-Distribution Samples

Notation

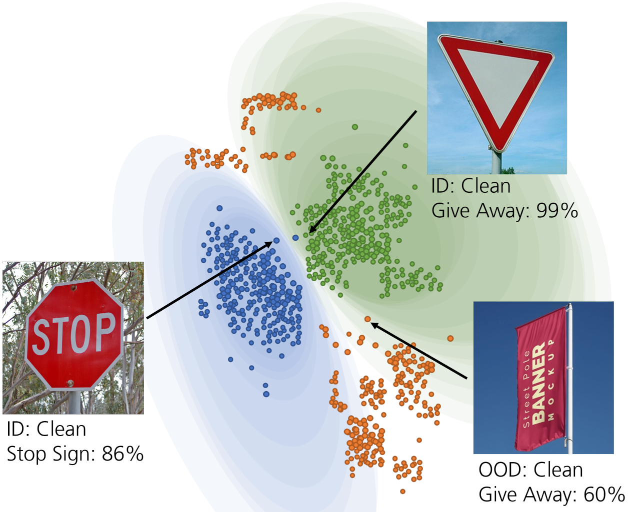
- Classifier $f(x)$
- Confidence, e.g. $h(x) = \text{SoftMax}(f(x))$
- In-distribution (ID) $x \sim \mathcal{D}_{in}$ (●, ●)
- Adversarial attack $\tilde{x}$ (ℓ_2-norm)



Background - Adversarial Attacks for In- and Out-of-Distribution Samples

Notation

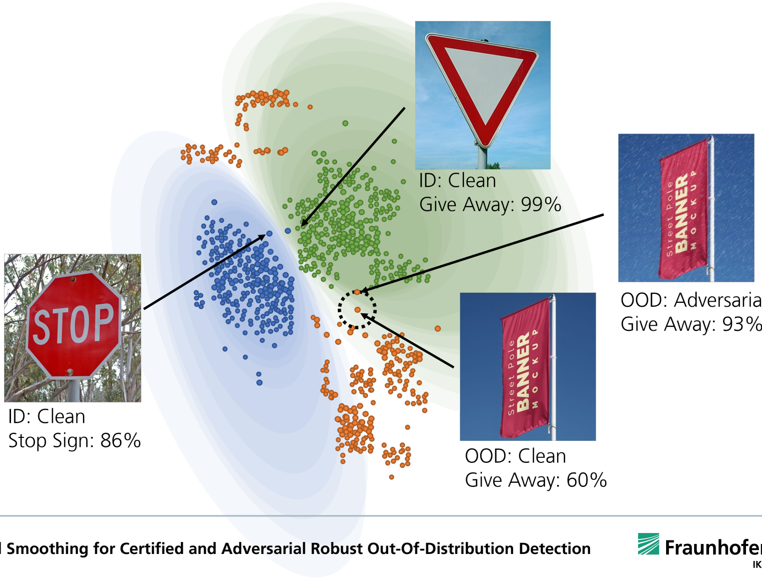
- Classifier $f(x)$
- Confidence, e.g. $h(x) = \mathrm{SoftMax}(f(x))$
- In-distribution (ID) $x \sim \mathcal{D}_{in}$ (●, ●)
- Adversarial attack $\tilde{x}$ (ℓ_2-norm)
- Out-Of-Distribution (OOD) $z \sim \mathcal{D}_{out}$ (●)



Background - Adversarial Attacks for In- and Out-of-Distribution Samples

Notation

- Classifier $f(x)$
- Confidence, e.g. $h(x) = \mathrm{SoftMax}(f(x))$
- In-distribution (ID) $x \sim \mathcal{D}_{in}$ (●, ●)
- Adversarial attack $\tilde{x}$ (ℓ_2-norm)
- Out-Of-Distribution (OOD) $z \sim \mathcal{D}_{out}$ (●)



Background - Certified Robustness with Randomized Smoothing

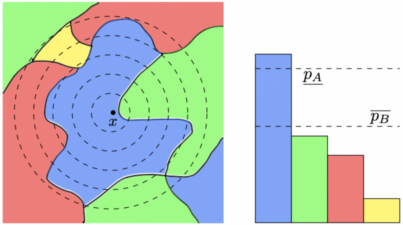


Figure: Randomized Smoothing [Cohen et al., 2019, Lecuyer et al., 2019]

Background - Certified Robustness with Randomized Smoothing

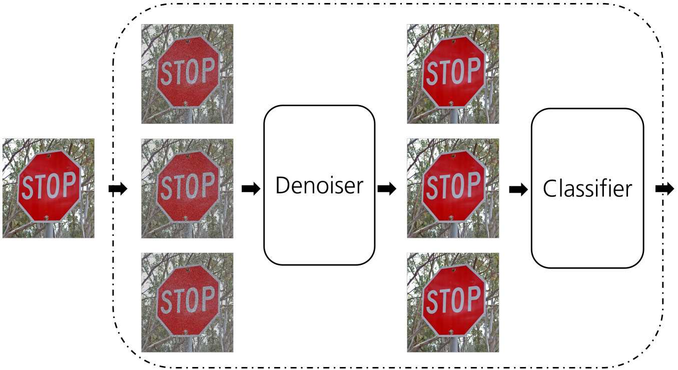


Figure: Diffusion Denoised Smoothing
[Salman et al., 2020, Carlini et al., 2023]

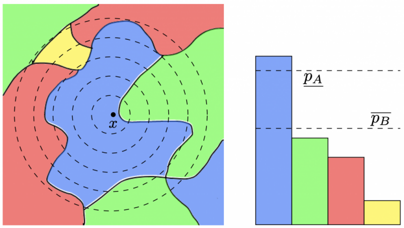


Figure: Randomized Smoothing
[Cohen et al., 2019,
Lecuyer et al., 2019]

Our Contribution - Certified Robustness for the Maximum Confidence

Theorem

Let $F : \mathbb{R}^d \to \mathbb{P}(\mathcal{Y})$ be any soft classifier and G be its associated smooth classifier defined as:

$$G(x) \stackrel{def}{=} \mathop{\mathbb{E}}_{\delta \sim \mathcal{N}(0,\sigma^2 I)} \left[F(x+\delta) \right],$$

with $\sigma > 0$. If $p = \max_{y \in \mathcal{Y}} G(x)_y > 1/2$, then, we have that:

$$\max_{y \in \mathcal{Y}} G(x+\delta)_y \leq \sqrt{\frac{2}{\pi}} \Phi^{-1}(p) + p,$$

for every $\|\delta\|_2 < \sigma \Phi^{-1}(p)$.

Our Contribution - Certified Robustness for the Maximum Confidence

Theorem

Let $F : \mathbb{R}^d \to \mathbb{P}(\mathcal{Y})$ be any soft classifier and G be its associated smooth classifier defined as:

$$G(x) \overset{def}{=} \underset{\delta \sim \mathcal{N}(0,\sigma^2 I)}{\mathbb{E}} \left[F(x+\delta) \right],$$

with $\sigma > 0$. If $p = \max_{y \in \mathcal{Y}} G(x)_y > 1/2$, then, we have that:

$$\max_{y \in \mathcal{Y}} G(x+\delta)_y \leq \sqrt{\frac{2}{\pi}} \Phi^{-1}(p) + p,$$

for every $\|\delta\|_2 < \sigma \Phi^{-1}(p)$.

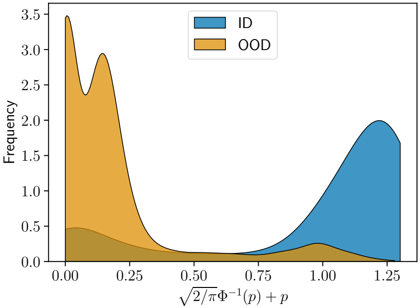


Figure: Distribution of the certified smooth ($\sigma = 0.12$) scores (*maximum confidence*) on ID (`CIFAR10`) and OOD (all other datasets) samples.

DISTRO: DIffusion denoised SmooThing for Robust OOD detection

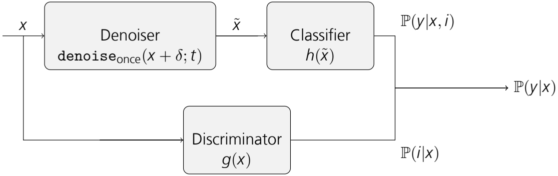


Figure: Overview of DISTRO

- $\mathbb{P}(y|x) = \mathbb{P}(y|x,i)\mathbb{P}(i|x) + \frac{1}{K}(1 - \mathbb{P}(i|x))$
- $\mathbb{P}(y|x,i) = h(\texttt{denoise}_{\text{once}}(x+\delta; t))$
- $\mathbb{P}(i|x) = \frac{1}{1+e^{-g(x)}}$

DISTRO: DIffusion denoised SmooThing for Robust OOD detection

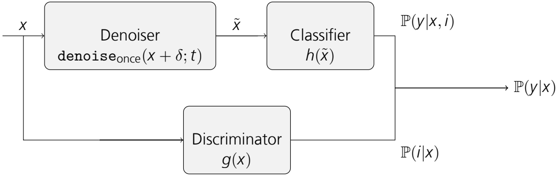


Figure: Overview of DISTRO

- $\mathbb{P}(y|x) = \mathbb{P}(y|x,i)\mathbb{P}(i|x) + \frac{1}{K}(1 - \mathbb{P}(i|x))$
- $\mathbb{P}(y|x,i) = h(\texttt{denoise}_{\text{once}}(x+\delta;t))$
- $\mathbb{P}(i|x) = \frac{1}{1+e^{-g(x)}}$

Asymptotic Confidence

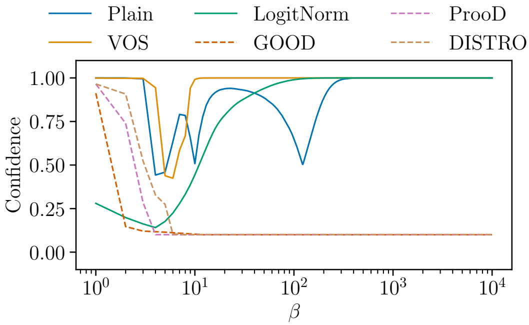


Figure: Two categories: *standard* (continuous line) and *guaranteed* (dashed line).

Comparison between this work and previous methods

Methods	In-Distribution (ID) Accuracy			Out-Of-Distribution (OOD) Detection				
	Clean	Adversarial ℓ_∞	Certified ℓ_2	Clean	Adversarial ℓ_∞	Certified ℓ_∞	Certified ℓ_2	Asymptotic underconfidence
- **Standard**								
OE [Hendrycks et al., 2019]	✓			✓				
VOS [Du et al., 2021]	✓			✓				
LogitNorm [Wei et al., 2022]	✓			✓				
- **Adversarial**								
ACET [Hein et al., 2019]	(✓)	✓		✓	(✓)			
ATOM [Chen et al., 2021]	(✓)			✓	(✓)			
- **Guaranteed**								
GOOD [Bitterwolf et al., 2020]					✓	✓		✓
ProoD [Meinke et al., 2022]	✓			✓	✓	✓		✓
DISTRO (Our)	✓	✓	✓	✓	✓	✓	✓	✓

In-Distribution Results

Adversarial Accuracy

- AutoAttack with ℓ_∞-norm attacks
- budget $\epsilon \in \{2/255, 8/255\}$

Certified Accuracy

- Randomized Smoothing
- 10'000 Gaussian distributed samples
- failure probability of 0.001
- All $R > 0$ are considered

Table: **ID Accuracy**: Results of clean, adversarial and certified accuracy (%) on the `CIFAR10` test set. The grayed-out models have an accuracy drop greater than 3% relative to the model with the highest accuracy.

Method	Clean	Adversarial (ℓ_∞) $\epsilon = 2/255$	Adversarial (ℓ_∞) $\epsilon = 8/255$	Certified (ℓ_2) $\sigma = 0.12$	Certified (ℓ_2) $\sigma = 0.25$
Plain*	95.01	2.16	0.00	28.14	14.17
OE*	95.53	1.97	0.00	31.48	10.88
VOS†	94.62	2.24	0.00	13.13	10.02
LogitNorm‡	94.48	2.65	0.00	12.53	10.25
ATOM*	92.33	0.00	0.00	0.00	0.00
ACET*	91.49	69.01	6.04	57.13	12.48
GOOD^*_{80}	90.13	11.65	0.23	17.33	10.31
ProoD* $\Delta = 3$	95.46	2.69	0.00	33.92	13.50
DDS	**95.55**	72.97	24.09	82.26	64.58
DISTRO (our)	95.47	**73.34**	**27.14**	**82.77**	**65.63**

Results for ID: CIFAR10

Table: **Robust OOD detection.** We consider the following metrics: clean top-1 accuracy on CIFAR10/100 test sets, clean AUC, guaranteed (GAUC), adversarial AUC (AAUC), clean AUPR, guaranteed AUPR (GAUPR), adverarial AUPR (AAUPR), clean FPR95% (FPR), guaranteed FPR95% (GFPR) and adversarial FPR95% (AFPR). Averaging was performed on a variety of OOD datasets. We consider MSP [Hendrycks and Gimpel, 2017] for all methods and metrics (with temperature $T = 1$). The guaranteed ℓ_2-norm is computed for $\sigma = 0.12$ for all $R > 0$, while the adversarial and guaranteed ℓ_∞-norm are computed for $\epsilon = 0.01$. The grayed-out models have an accuracy drop greater than 3% relative to the model with the highest accuracy. **Bold** numbers are superior results.

ID: CIFAR10	Acc.	AUC↑	GAUC↑		AAUC↑	AUPR↑	GAUPR↑		AAUPR↑	FPR↓	GFPR↓		AFPR↓
			ℓ_2	ℓ_∞	ℓ_∞		ℓ_2	ℓ_∞	ℓ_∞		ℓ_2	ℓ_∞	ℓ_∞
- **Standard**													
Plain*	95.01	94.56	48.86	0.00	24.52	99.42	60.05	0.00	82.30	35.72	100.0	100.0	96.72
OE*	**95.53**	**98.78**	46.88	0.00	37.91	**99.87**	63.08	0.00	84.49	**4.71**	100.0	100.0	70.26
VOS†	94.62	90.82	30.13	0.00	20.62	99.15	41.62	0.00	81.80	61.66	**94.10**	100.0	100.0
LogitNorm‡	94.48	96.71	40.73	0.00	39.76	99.64	49.31	0.00	86.47	13.95	100.0	100.0	91.10
- **Adversarial**													
ACET*	91.48	97.24	60.21	0.00	93.01	99.68	76.22	0.00	99.16	13.82	95.65	100.0	32.15
ATOM*	92.33	98.82	97.15	0.00	44.65	99.86	95.51	0.00	85.74	4.14	5.04	100.0	62.65
- **Guaranteed**													
GOOD^*_{80}	90.13	93.12	36.45	57.52	78.11	99.22	52.31	89.54	95.19	30.00	100.0	72.45	47.55
ProoD* $\Delta = 3$	95.46	98.72	52.36	**59.56**	64.22	**99.87**	66.53	**93.89**	94.52	5.49	100.0	100.0	86.49
DISTRO (our)	95.47	98.71	**53.37**	59.49	**89.36**	**99.87**	**68.45**	93.88	**98.70**	5.44	100.0	100.0	**51.15**

Results for ID: CIFAR100

Table: **Robust OOD detection.** We consider the following metrics: clean top-1 accuracy on CIFAR10/100 test sets, clean AUC, guaranteed (GAUC), adversarial AUC (AAUC), clean AUPR, guaranteed AUPR (GAUPR), adverarial AUPR (AAUPR), clean FPR95% (FPR), guaranteed FPR95% (GFPR) and adversarial FPR95% (AFPR). Averaging was performed on a variety of OOD datasets. We consider MSP [Hendrycks and Gimpel, 2017] for all methods and metrics (with temperature $T = 1$). The guaranteed ℓ_2-norm is computed for $\sigma = 0.12$ for all $R > 0$, while the adversarial and guaranteed ℓ_∞-norm are computed for $\epsilon = 0.01$. The grayed-out models have an accuracy drop greater than 3% relative to the model with the highest accuracy. **Bold** numbers are superior results.

ID: CIFAR100	Acc.	AUC↑	GAUC↑		AAUC↑	AUPR↑	GAUPR↑		AAUPR↑	FPR↓	GFPR↓		AFPR↓
			ℓ_2	ℓ_∞	ℓ_∞		ℓ_2	ℓ_∞	ℓ_∞		ℓ_2	ℓ_∞	ℓ_∞
- **Standard**													
Plain*	**77.38**	81.60	30.63	0.00	16.98	97.84	45.10	0.00	81.27	82.52	100.0	100.0	100.0
OE*	77.28	90.41	39.87	0.00	22.79	98.90	49.46	0.00	81.96	47.49	100.0	100.0	87.74
- **Adversarial**													
ACET*	74.47	90.27	36.36	0.00	27.68	98.84	43.50	0.00	82.60	44.11	**90.41**	100.0	**74.99**
ATOM*	71.73	91.72	84.38	0.00	31.52	98.88	79.95	0.00	83.36	30.81	30.09	100.0	73.69
- **Guaranteed**													
ProoD* $\Delta = 1$	76.79	**90.90**	42.83	**37.67**	43.81	**98.91**	50.90	**89.66**	90.46	42.12	100.0	100.0	97.11
DISTRO (our)	76.83	90.89	**47.74**	37.53	**65.16**	98.90	**55.26**	89.63	**94.78**	**40.94**	100.0	100.0	87.81

Conclusion

Table: Overall average between the metrics for `CIFAR10/100`.

Method	Average	
	C-10	C-100
Plain	44.02	34.48
OE	50.12	40.42
VOS	38.60	-
LogitNorm	46.31	-
ACET	59.64	41.86
ATOM	64.79	54.38
GOOD_{80}	64.74	-
ProoD $\Delta = 3$	64.09	52.51
DISTRO (our)	**77.08**	**59.95**

- Surprisingly, ATOM shows similar results as ProoD and GOOD.
- It is evident that the ℓ_2-norm GAUC (and GAUPR) diverge from zero when standard OOD detection models are considered.

Code on Github

References I

Bitterwolf, J., Meinke, A., and Hein, M. (2020).
Certifiably adversarially robust detection of out-of-distribution data.
In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Carlini, N., Tramer, F., Zico Kolter, J., et al. (2023).
(certified!!) adversarial robustness for free!
In *Submitted to The Eleventh International Conference on Learning Representations*.
under review.

Chen, J., Li, Y., Wu, X., Liang, Y., and Jha, S. (2021).
Atom: Robustifying out-of-distribution detection using outlier mining.
In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 430–445. Springer.

Cohen, J., Rosenfeld, E., and Kolter, Z. (2019).
Certified adversarial robustness via randomized smoothing.
In *International Conference on Machine Learning*, pages 1310–1320. PMLR.

Du, X., Wang, Z., Cai, M., and Li, Y. (2021).
Vos: Learning what you don't know by virtual outlier synthesis.
In *International Conference on Learning Representations*.

References II

Hein, M., Andriushchenko, M., and Bitterwolf, J. (2019).
Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem.
In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 41–50. Computer Vision Foundation / IEEE.

Hendrycks, D. and Gimpel, K. (2017).
A baseline for detecting misclassified and out-of-distribution examples in neural networks.
In *International Conference on Learning Representations*.

Hendrycks, D., Mazeika, M., and Dietterich, T. G. (2019).
Deep anomaly detection with outlier exposure.
In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Lecuyer, M., Atlidakis, V., Geambasu, R., Hsu, D., and Jana, S. (2019).
Certified robustness to adversarial examples with differential privacy.
In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 656–672. IEEE.

Meinke, A., Bitterwolf, J., and Hein, M. (2022).
Provably robust detection of out-of-distribution data (almost) for free.
In *NeurIPS*.

References III

Salman, H., Sun, M., Yang, G., Kapoor, A., and Kolter, J. Z. (2020).
Denoised smoothing: A provable defense for pretrained classifiers.
Advances in Neural Information Processing Systems, 33:21945–21957.

Wei, H., Xie, R., Cheng, H., Feng, L., An, B., and Li, Y. (2022).
Mitigating neural network overconfidence with logit normalization.
In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S., editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23631–23644. PMLR.