Provably Safeguarding a Classifier from OOD and Adversarial Samples

ICLR 2025

Nicolas Atienza (Thales cortAIx Labs, LISN), Christophe Labreuche (Thales cortAIx Labs), Johanne Cohen (LISN), Michèle Sebag (LISN)

April 2025

Motivation

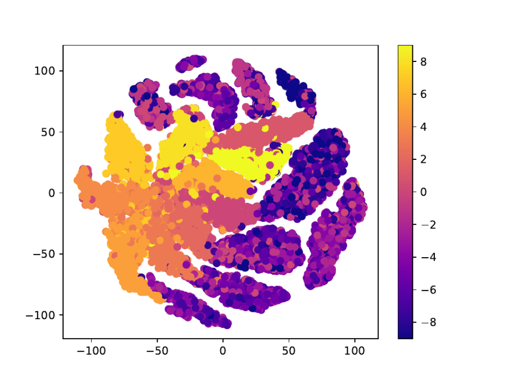
Robust AI
Towards reliable ML models that can handle real-world distributions Ye et al. 21

SPADE: Sample-efficient ProbAbilistic Detection using Extreme Value Theory

- Transform a trained classifier into a trustful abstaining classifier
- Detect OOD samples and adversarial attacks with provable guarantees

SPADE Overview

Observation: OOD samples are not mixed with in-distribution samples within latent space.



- Cifar-10 classes
- FGSM attacks

SPADE Process: Use distance to nearest neighbors from ID as proxy to detect OOD samples.

Lee et al. 18, Sun et al. 22

SPADE Overview

Distance Based SOTA Limitations

- Distance-based methods are empirically explored — Yang et al. 24
- Strong assumptions on the ID (and OOD !) latent data distribution — Lee et al. 18
- Empirical analysis with high computational cost — Sun et al. 22

SPADE Contributions

Formal analysis of the z_c behavior (latent distance to nearest ID neighbors)

Pros

- Formal probabilistic Guarantees
- No requirement on ID/OOD distribution
- Computationally frugal and stable

Extreme Value Distribution

Fisher-Tippett 28

Let Z be a random variable over the real-valued space $\mathbb{R}$. Let $Z^{(\ell)}$ be the random variable defined as the normalised maximum value over ℓ independent drawings of Z. When ℓ goes to infinity, the *limiting distribution* of $Z^{(\ell)}$ is the cumulative distribution $P(Z^{(\ell)} < z) \underset{\ell \to \infty}{\to} G_{\xi,\mu,\sigma}(z)$, expressed as one of the two parametric models:

$$G_{\xi,\mu,\sigma}(z) = \exp \left\{ \begin{array}{ll} \left(1 + \xi \frac{z-\mu}{\sigma}\right)_+^{-1/\xi} & \text{if } \xi \neq 0 \\ -\exp\left(\frac{\mu - z}{\sigma}\right) & \text{otherwise} \end{array} \right\} \tag{1}$$

with $\mu \in \mathbb{R}$ a location parameter, $\sigma \in \mathbb{R}_+$ a dispersion parameter and $\xi \in \mathbb{R}$ a shape parameter referred to as *extreme value index*.

$\rightarrow$ EVT has been applied to anomaly detection tasks

Goix et al. 16, Siffer et al. 17, French et al. 19

SPADE Overview : OOD Test Based on Extreme Value Theory

Definition: OOD Test

The OOD Test is defined as the minimum probability among classes of the distance to a nearest neighbors being extreme

$$OOD(x) = \min_{c \in \mathcal{Y}} G^{(c)}(z_c)$$

where $G^{(c)}(z_c)$ is the probability of latent distance z_c to ID nearest neighbors of class c to be extreme.

Definition: Abstaining Classifier

classifier f_τ abstains from making predictions on a sample x if x is considered to be extreme with probability at least $1 - \tau$ w.r.t. its candidate class $c = f(x)$.

$$f_\tau(x) = \begin{cases} f(x) & \text{if } z_c \leq {G^{(c)}}^{-1}(1-\tau) \\ \text{abstain} & \text{otherwise} \end{cases}$$

SPADE Overview : Adversarial Robust Bound

Theorem

Assume that the latent embedding h is K-Lipschitz. Let x be an adversarial sample obtained by perturbing a training sample x^* of class c, with perturbation amplitude $\varepsilon = \|x - x^*\|$, and let $f(x) = c' \neq c$. Let x'^* of class c' denote the k-th nearest training sample in $\mathcal{D}$ of x.

Then, with probability at least $1 - \tau$ either f_τ abstains on x, or ε admits the following lower bound:

$$\varepsilon \geq \frac{1}{K}\left(G^{(c,c')^{-1}}(1-\tau) - G^{(c)^{-1}}(1-\tau)\right)$$

Limitation

What if Lipschitz constant very large ?

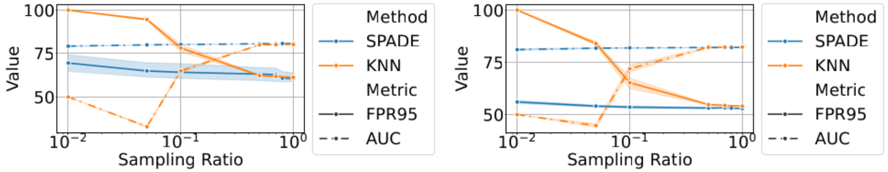
$\rightarrow$ Extension Thm with local Lipschitz constant.

SPADE OOD Detection Performance

	Near OOD				Far OOD						
	SSB Hard		NINCO		iNaturalist		Textures		OpenImages-O		Rank
	AUC ↑	FPR95 ↓	AUC ↑	FPR95 ↓	AUC ↑	FPR95 ↓	AUC ↑	FPR95 ↓	AUC ↑	FPR95 ↓	
MSP (1)	**72.53**	**74.43**	**80.66**	**57.72**	87.78	44.08	82.81	59.16	85.21	49.62	**3**
ODIN (2)	72.51	77.36	77.55	70.83	**89.51**	**41.46**	87.02	56.58	86.33	54.10	**4**
MDS (3)	52.15	90.46	68.49	71.66	76.49	56.07	94.11	27.07	77.68	59.66	**5**
KNN (4)	62.80	84.08	79.30	58.92	84.62	42.39	**96.06**	**23.39**	**86.38**	**44.24**	**1**
SPADE	61.91	85.27	77.99	61.04	85.26	44.84	95.86	24.63	85.79	46.33	**2**

→ SPADE is robust second w.r.t. SOTA methods on CIFAR-10, CIFAR-100, ImageNet on both standard OOD detection tasks and adversarial detection (FGSM, PGD, Auto-Attack).

SPADE OOD Detection Performance



(a) near OOD (b) far OOD

$\rightarrow$ Good performance w.r.t. aggressively subsampling the training set.

SPADE Conclusion and Perspectives

Strengths

- Detection performance on-par with SOTA models
- Low sample complexity
- Probabilistic formal guarantees

Limitations

- Dependence wrt Lipschitz constant alleviated based on the use of local Lipschitz constants

Perspective:

- Extending training loss to better support robustness guarantees.

Thanks for your attention

Feel free to reach out to discuss !

We’ll be delighted to discuss

(we are hiring)

christophe.labreuche@thalesgroup.com https://github.com/natixx14/SPADE https://arxiv.org/abs/2501.10202