PROVABLY SAFEGUARDING A CLASSIFIER FROM OOD AND ADVERSARIAL SAMPLES: AN EXTREME VALUE THEORY APPROACH

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper introduces a novel method, *Sample-efficient Probabilistic Detection using Extreme Value Theory* (SPADE), which transforms a classifier into an abstaining classifier, offering provable protection against out-of-distribution and adversarial samples. The approach is based on a Generalized Extreme Value (GEV) model of the training distribution in the classifier's latent space, enabling the formal characterization of OOD samples. Interestingly, under mild assumptions, the GEV model also allows for formally characterizing adversarial samples. The abstaining classifier, which rejects samples based on their assessment by the GEV model, provably avoids OOD and adversarial samples. The empirical validation of the approach, conducted on various neural architectures (ResNet, VGG, and Vision Transformer) and medium and large-sized datasets (CIFAR-10, CIFAR-100, and ImageNet), demonstrates its frugality, stability, and efficiency compared to the state of the art.

1 INTRODUCTION

A key challenge in deploying learned models in real-world settings is managing out-of-distribution (OOD) samples. When a learned model encounters data that deviates from the training distribution, it can lead to failures with significant consequences, particularly in high-stakes applications such as medical diagnosis, autonomous driving or risk analysis (Salehi et al., 2022; Yang et al., 2022). The ultimate aim in machine learning is to achieve OOD generalization, where the model encapsulates the core concept with sufficient accuracy to effectively handle atypical but real samples (Ye et al., 2021). A step towards OOD generalization is OOD detection, which equips the learned model with the ability to recognize atypical samples and refrain from making risky predictions. OOD detection is approached from several directions, including methods based on classification (Hendrycks & Gimpel, 2017; Liang et al., 2018; Hsu et al., 2020), reconstruction metrics (Jiang et al., 2023; Li et al., 2023), density estimation (Ren et al., 2019; Liu et al., 2021; Du et al., 2022b) and distance-based estimation (Papernot & McDaniel, 2018; Sun et al., 2022; Dziedzic et al., 2022; Ming et al., 2023) (more in Section 6).

OOD detection is complicated by the fact that, to the best of our knowledge, no universally accepted definition exists for what qualifies as an OOD sample. The boundaries between in-distribution and OOD data are inherently ambiguous and different domain experts may classify the same sample differently based on their understanding and experience (Idrissi et al., 2023). Consequently, the validation of OOD detection methods relies heavily on experimental studies using well-curated datasets, such as near and far OOD datasets (Yang et al., 2022).

It is worth noting that human experts and models tend to make different decisions regarding both OOD and adversarial samples (Goodfellow et al., 2015), albeit in distinct ways. Experts typically recognize an OOD sample as belonging to a given class, despite its atypicality, while the model assigns it to a random class. Conversely, experts perceive an adversarial example as typical of a specific class, yet the model confidently misclassifies it into a different class.

The approach presented in this paper, referred to as *Sample-efficient ProbAbilistic Detection using Extreme value theory* (SPADE) and inspired by distance-based approaches (Sun et al., 2022), introduces an original model of the training distribution relative to a learned model (hereafter the

teacher). Specifically, the distances between samples in the teacher's latent space are modelled using the Extreme Value Theory (EVT) (Fisher & Tippett, 1928). This model provides a sound and robust test for detecting and rejecting OOD samples. Most interestingly, under mild assumptions this test also provably rejects adversarial examples with high probability, subject to a bound on their perturbation amplitude.

The contributions of the proposed approach are fourfold: i) it introduces a formal definition of OOD samples relative to a teacher and its latent representation; ii) this definition leads to a statistically frugal OOD test based on EVT first principles; iii) this test operationally rejects OOD and adversarial samples with provable guarantees; iv) the effectiveness of the approach is experimentally and successfully demonstrated against strong baselines for learned models with different architectures (He et al., 2016; Dosovitskiy et al., 2021; Ming et al., 2023).

The paper is organized as follows. Section 2 outlines the formal background of distance-based OOD detection and introduces extreme value theory. Section 3 gives an overview of the SPADE approach and its formal analysis. Sections 4 and 5 respectively detail the experimental setup and the experiments conducted to validate SPADE against state-of-the-art methods. Section 6 discusses the contributions within the context of related work, and the paper concludes with perspectives for future research.

Notations. Let $\mathcal{D} = \{(\mathbf{x}_1, y_1) \dots (\mathbf{x}_n, y_n)\}$ denote the training set, iid drawn after the joint distribution $P_{X,Y}$, with $\mathcal{X}$ the instance space and $\mathcal{Y} = \{1, \dots n_c\}$ the set of classes. The trained teacher f is expressed as $f = g \circ h$, where h is the embedding in the latent space ($h : \mathcal{X} \mapsto \mathbb{R}^d$) and g is the mapping used to make decisions based on the latent representation.

2 Formal Background

This section describes the main concepts defined by (Ye et al., 2021) in the context of distance-based OOD generalization, and briefly introduces extreme value theory for completeness.

2.1 Properties of Latent In-Distribution

As mentioned above, the complexity of OOD characterization corresponds to the highly complex and diverse nature of real-world data (Farquhar & Gal, 2022). In the literature (Cimpoi et al., 2014; Horn et al., 2018; Wang et al., 2022; Vaze et al., 2022; Bitterwolf et al., 2023) OOD characterization often involves subjective assessment ("data that appear noticeably different from in-distribution to human observers"). A common definition states that the OOD is different from the known in-distribution, e.g., the training distribution, although this definition does not capture the specifics of OOD: OOD is not merely any distribution that differs from ID.

On the other hand, as noted by Papernot & McDaniel (2018); Sun et al. (2022), the representation in latent space of the in-distribution (hereafter ID) presents distinct characteristics, such as being formed of compact and well-separated clusters. Ye et al. (2021) formalize these properties in terms of variation and informativeness of the latent representation:

Definition 1 ((Ye et al., 2021))**.** *The* variation of embedding h *across a finite distribution* $\mathcal{D}$*, noted* $\mathcal{V}_\rho(h, \mathcal{D})$*, is defined as the maximum diameter over all classes* c *of the ball containing distribution* $h(\mathbf{x})$ *for* $(\mathbf{x}, y) \in \mathcal{D}$ *and* $y = c$*:*

$$\mathcal{V}(h, \mathcal{D}) = \max_{c \in Y} \sup_{\substack{(\mathbf{x},c)\in\mathcal{D} \\ (\mathbf{x}',c)\in\mathcal{D}}} \|h(\mathbf{x}) - h(\mathbf{x}')\| \tag{1}$$

where the distance $\|h(\mathbf{x}) - h(\mathbf{x}')\|$ *is usually set to* L_2 *distance*[1]*. Embedding* h *is said* η*-invariant across* $\mathcal{D}$ *if* $\mathcal{V}(h, \mathcal{D}) < \eta$*.*

The variation of embedding h thus measures the maximum thickness and width of the latent manifold containing the image of (samples in) a class.

[1]The Kullback Leibler divergence is considered when embedding h is a probabilistic one. See Ye et al. (2021) for more detail.

Definition 2 ((Ye et al., 2021))**.** *The* informativeness of embedding h *across a finite distribution* $\mathcal{D}$*, noted* $\mathcal{I}(h, \mathcal{D})$*, is defined as the average over all pairs* (c, c') *of distinct classes, of the minimum distance between* $h(\mathbf{x})$ *and* $h(\mathbf{x}')$ *for* $(\mathbf{x}, y)$ *and* $(\mathbf{x}', y')$ *in* $\mathcal{D}$*, with* $y = c \neq y' = c'$*:*

$$\mathcal{I}(h, \mathcal{D}) = \frac{1}{n_c(n_c - 1)} \sum_{c \neq c' \in \mathcal{Y}} \min_{\substack{(\mathbf{x},y) \in \mathcal{D}, (\mathbf{x}',y') \in \mathcal{D} \\ y = c \neq y' = c'}} \|h(\mathbf{x}) - h(\mathbf{x}')\| \tag{2}$$

where the distance is usually set to L_2 *distance (see footnote 1), and* n_c *denotes the number of classes. The latent embedding* h *is said* δ-informative *across* $\mathcal{D}$ *if* $\mathcal{I}(h, \mathcal{D}) > \delta$*.*

It is noted that informativeness and variation are closely related to the *compactness* and *dispersion* metrics introduced and optimized in CIDER to achieve OOD generalization (Ming et al., 2023).

These definitions are defined in terms of supremum over classes, raising the challenge of their statistical stability and algorithmic exploitation (e.g., through setting thresholds). The approach presented here addresses this challenge by exploiting extreme value theory, as described below and referring the reader to Haan & Ferreira (2006) for a comprehensive introduction.

2.2 Extreme Value Theory

Dating back to Fisher & Tippett (1928); Gnedenko (1943), Extreme Value Theory (EVT) focuses on modeling and understanding the tail behavior of distributions. EVT is based on the premise that, under mild assumptions, the distributions of extreme events converge to a common form, even if their original distributions differ. For instance, while the distributions of seismic intensities and the heights of rogue waves − factors that respectively influence the design of buildings and oil rigs − may differ, their extreme values (maxima) are governed by the same class of distributions. This limiting distribution is known as the *Extreme Value Distribution* (EVD):

Definition 3 (Extreme Value Distribution (EVD) (Fisher & Tippett, 1928))**.** *Let* Z *be a random variable over the real-valued space* $\mathbb{R}$*. Let* $Z^{(\ell)}$ *denote the random variable defined as the 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:*

$$G_{\xi,\mu,\sigma}(z) = \exp \left\{ \begin{array}{ll} \left(1 + \xi \frac{z-\mu}{\sigma}\right)_+^{-1/\xi} & \textit{if } \xi \neq 0 \\ -\exp\left(\frac{\mu - z}{\sigma}\right) & \textit{otherwise} \end{array} \right\} \tag{3}$$

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.

Overall, the EVT framework provides a general parametric model for extreme events associated with a random variable, independent of the distribution of Z itself. The universality of these models reflects the fact that modeling the extreme events associated with a distribution relies only on the behavior of its tail. This tail can take one of three forms: (i) an exponential tail ($\xi = 0$, corresponding to the Gumbel distribution); (ii) a heavy tail ($\xi > 0$, corresponding to the Fréchet distribution); or (iii) a bounded tail ($\xi < 0$, corresponding to the Weibull distribution). Despite its applicability, EVT has, to the best of our knowledge, seen limited use in machine learning, with notable exceptions in the area of anomaly detection (Smith et al., 2012; Siffer et al., 2017; French et al., 2019).

3 SPADE Overview

Aimed at OOD detection, SPADE proposes a formal characterization of OOD concerning a trained teacher model and the associated latent representation on the one hand and the training distribution (hereafter in-distribution, ID) on the other hand. This characterization relies on generalized extreme value (GEV) models, which allow for detecting and rejecting out-of-ID samples. For simplicity and by abuse of language, out-of-ID samples are referred to as OOD in the following. Interestingly, under mild assumptions, the GEV models also allow for detecting adversarial samples. The abstaining classifier, equipped with the GEV-based detection tests, provides probabilistic guarantees of OOD and adversarial sample rejection, subject to a lower bound on the magnitude of the adversarial perturbation.

3.1 EVT-BASED CHARACTERIZATION OF OOD

The distance-based OOD detection literature (see e.g., (Papernot & McDaniel, 2018; Dziedzic et al., 2022; Sun et al., 2022)) suggests that a sample is *likely* to be an OOD sample if it is *sufficiently* distant from the training samples of all (or most) classes in the latent space.

In SPADE, this process is reformulated using generalized extreme value models, directly yielding the probability for a sample to be OOD. Let (X, Y) denote the random variable following the joint distribution $P_{X,Y}$. For $Y = c$, let Z_c be the random variable defined as the distance between $h(X)$ and its k-th nearest neighbor in latent distance, belonging to $\mathcal{D}$ with same class c. By definition, the limiting distribution of the maxima of Z_c follows a Generalized Extreme Value model noted $G^{(c)}$, with $Pr(Z_c^{(l)} < v) \underset{\ell \to \infty}{\to} G^{(c)}(v)$.

For each sample $\mathbf{x}$ in the instance space and each class c, let z_c be defined as $\|h(\mathbf{x}) - h(\mathbf{x}_{knn,c})\|$ with $\mathbf{x}_{knn,c}$ the k-th nearest neighbor of $\mathbf{x}$ in latent space, such that $(x_{knn,c}, c)$ belongs to $\mathcal{D}$. As the true label of $\mathbf{x}$ is unknown at inference time, the proposed OOD test retains the lowest probability of $\mathbf{x}$ being OOD according to all $G^{(c)}$:

Definition 4 (OOD test)**.** *Let $\mathbf{x}$ denote an instance in $\mathcal{X}$ with z_c its Euclidean latent distance to its k-nearest neighbor of class c in the training set $\mathcal{D}$. The probability of $\mathbf{x}$ being an OOD sample, noted $OOD(\mathbf{x})$, is defined as:*

$$OOD(\mathbf{x}) = \min_{c \in \mathcal{Y}} G^{(c)}(z_c) \tag{4}$$

In other words, $\mathbf{x}$ is considered to be OOD if it is extreme concerning all GEV models associated with the different classes.

The decision to consider a separate GEV $G^{(c)}$ for each class c is intended to address situations where classes exhibit different levels of variation in the latent space. In such cases, considering a single GEV model for all classes could result in either erroneously rejecting samples from a class with high variation or incorrectly accepting OOD samples from a class with low variation.

Since the OOD test depends on the classifier's latent representation, one might wonder to what extent different tests based on different classifiers are consistent (as discussed further in Section 5).

3.2 ABSTAINING CLASSIFIER ON OOD SAMPLES

A classifier abstaining on OOD samples is built as follows.

Definition 5 (Abstaining classifier)**.** *Given teacher f and confidence $1 - \tau$, with $0 < \tau < 1$, classifier f_τ abstains from making predictions on a sample $\mathbf{x}$ if $\mathbf{x}$ is considered to be extreme with probability at least $1 - \tau$ w.r.t. its candidate class $c = f(\mathbf{x})$. With same notations as above:*

$$f_\tau(\mathbf{x}) = \begin{cases} f(\mathbf{x}) & \text{if } z_c \leq {G^{(c)}}^{-1}(1-\tau) \\ \textit{abstain} & \textit{otherwise} \end{cases} \tag{5}$$

where z_c is the distance between $h(\mathbf{x})$ and its nearest neighbor of class c in $\mathcal{D}$.

The abstention test embedded in f_τ is more precise than the OOD test (Def. 4) because it incorporates the additional information of $f(\mathbf{x})$. Note, however, that both tests coincide under the common assumption that an OOD sample is closer to the samples of its own class, all else being equal.

3.3 ABSTAINING CLASSIFIER WITH PROVABLE GUARANTEES W.R.T. ADVERSARIAL EXAMPLES

Let us consider the adversarial example $\mathbf{x}$ built by perturbing a training sample noted $\mathbf{x}^*$ of class c, and let $c' = f(\mathbf{x}) \neq c$ be its (wrong) class according to f. Under mild assumptions, it is shown that the abstaining classifier f_τ abstains on adversarial sample $\mathbf{x}$ with probability $1 - \tau$, subject to a lower bound on its perturbation amplitude.

Let $G^{(c,c')}$ denote the generalized extreme value model associated with the *minimum* latent distance among pairs of examples $(\mathbf{x}, \mathbf{x}')$ respectively belonging to class c and c':

$$G^{(c,c')}(v) = Pr\Big(\|h(X) - h(X')\| > v \;\Big|\; (X,Y) \sim P_{X,Y},\ (X',Y') \sim P_{X,Y},\ Y = c,\ Y' = c'\Big)$$

Theorem 1. *Let us assume that the latent embedding h is K-Lipschitz. Let $\mathbf{x}$ be an adversarial sample built by perturbation of a training sample $\mathbf{x}^*$ of class c, with perturbation amplitude ε ($\|\mathbf{x} - \mathbf{x}^*\| < \varepsilon$), and let $f(\mathbf{x}) = c' \neq c$. Let $\mathbf{x}'^*$ of class c' denote the k-th nearest training sample in $\mathcal{D}$ of $\mathbf{x}$. Then, with probability at least $1 - \tau$ either f_ϵ abstains on $\mathbf{x}$, or perturbation ϵ is greater than the following lower bound:*

$$\varepsilon \geq \frac{1}{K}\left(G^{(c,c')^{-1}}(1-\tau) - G^{(c)^{-1}}(1-\tau)\right) \tag{6}$$

Proof. If $\|h(\mathbf{x}) - h(\mathbf{x}'^*)\| > (G^{(c)})^{-1}(1-\tau)$, then f_τ abstains on $\mathbf{x}$ (Def. 5). Otherwise,

$$\begin{aligned}\|h(\mathbf{x}^*) - h(\mathbf{x}'^*)\| &\leq \|h(\mathbf{x}^*) - h(\mathbf{x})\| + \|h(\mathbf{x}) - h(\mathbf{x}'^*)\| \\ &\leq K\varepsilon + (G^{(c)})^{-1}(1-\tau)\end{aligned} \tag{7}$$

Moreover, with probability $1 - \tau$,

$$(G^{(c,c')})^{-1}(1-\tau) \leq \|h(\mathbf{x}^*) - h(\mathbf{x}'^*)\| \tag{8}$$

Putting together Eqs. 7 and 8 concludes the proof. □

3.4 Estimating the GEV Models

The estimation of the GEV models involved in SPADE is detailed in the case of $G^{(c)}$, defining its approximation $\widehat{G}^{(c)}$. The same process is used to learn an approximation of the $G^{(c,c')}$ models, with the only difference being that the considered extreme values are minima instead of maxima.

After Siffer et al. (2017), a straightforward approximation of an EVD proceeds by sampling the extreme events along the block maxima method. This fitting process is, however, found to be sensitive to the number of blocks and the block size. The proposed approach thus is the *Peak Over Threshold* (POT) method, which relies on the Pickands-Balkema-de-Haan theorem, often referred to as the *second theorem of EVT* (Balkema & de Haan, 1974; Pickands, 1975). Formally, the POT considers the occurrences of events bypassing a threshold t, noting that the distribution of these occurrences follows a Generalized Pareto Distribution (GPD):

$$F_{\xi,\mu,\sigma}(z) = Pr(Z - t > z | Z > t) = \begin{cases} 1 - \exp\left(-\frac{z-\mu}{\sigma}\right) & \text{if } \xi = 0 \\ 1 - \left(1 + \frac{\xi(z-\mu)}{\sigma}\right)^{-1/\xi} & \text{otherwise} \end{cases} \tag{9}$$

Informally, for each class c, POT proceeds by fitting a GPD model to samples z_c that are over a threshold t_c. Formally, for each sample $(\mathbf{x}_i, c)$ in $\mathcal{D}$, let $z_i \in \mathbb{R}$ be the distance of $\mathbf{x}_i$ to its k-th nearest neighbor belonging to class c, in latent space. Let $\mathcal{D}_c$ be the set including all such z_i for $z_i > t_c$.

The parameters (μ_c, σ_c, ξ_c) of model $\widehat{G}^{(c)}$ are learned by maximum likelihood estimation (MLE) on $\mathcal{D}_c$:

$$(\mu_c, \sigma_c, \xi_c) = \underset{\mu,\sigma,\xi}{\arg\max} \sum_{z \in \mathcal{D}_c} \mathcal{L}_{\mu,\sigma,\xi}(z) \tag{10}$$

with $\mathcal{L}$ the log-probability density function of the GPD. MLE is preferred over alternative methods, e.g. the method of moments, due to its higher robustness and efficiency (Siffer et al., 2017).

Eventually, model $\widehat{G}^{(c)}$ approximately characterizes whether a given sample is OOD with respect to class c with confidence $1 - \tau$. The OOD test (Def. 4) is accordingly approximated as:

$$\widehat{OOD}(\mathbf{x}) = \min_{c \in \mathcal{Y}} \widehat{G}^{(c)}(z_c) \tag{11}$$

3.5 Discussion

SPADE retains the main benefits of distance-based OOD detection approaches, being agnostic to the structure of the OOD distribution and easy to implement. Furthermore, being grounded in the

EVT first principles, it enables estimating the probability for a sample to be OOD. Lastly, the abstaining classifier f_τ can also reject adversarial samples, subject to a lower bound on the magnitude of adversarial perturbations.

A potential weakness is the complexity of approximating GEV models, which is quadratic with respect to the number n of samples. This raises the question of whether stable and accurate GEV approximations can be achieved when aggressively subsampling the training set. A second issue pertains to the effectiveness of the lower bound used in rejecting adversarial samples (Eq. 6); specifically the question is whether $G^{(c,c')^{-1}}(1-\tau) - G^{(c)^{-1}}(1-\tau)$ is strictly positive for non-trivial confidence levels $1-\tau$.[2]

4 Experimental Setting

This section outlines the experimental setup used to evaluate SPADE in comparison to the state of the art. All experiments were conducted on Tesla A100 80GB GPUs. Further details are provided in the supplementary material (SM).

Goals. The experiments aim to empirically address four questions: the performance of the OOD detection tests based on the GEV models, particularly in comparison with distance-based approaches (Q1); their sensitivity with respect to the considered teacher (Q2); similarly, the performance of the adversarial sample detection test based on the GEV models, compared with state-of-the-art methods (Q3); the computational complexity and stability of the approximate GEV models embedded in SPADE (Q4).

The performance of SPADE is compared to five established baselines: the seminal MSP (Hendrycks & Gimpel, 2017), ODIN (Liang et al., 2018), MDS (Lee et al., 2018), KNN (Sun et al., 2022), and its extension CIDER (Ming et al., 2023) (further discussed in Section 6).

Metrics. The comparative evaluation is conducted using the OpenOOD framework (Zhang et al., 2023), with performance assessed by standard indicators: The *Area Under the ROC Curve (AUC)* measures the average rate of correct OOD/adversarial sample detection across all confidence levels $1-\tau$, corresponding to the true positive rate, as described in (Zhang et al., 2023). The *FPR95 indicator* represents the fraction of true samples misclassified as OOD (respectively, adversarial) when the detection threshold ensures 95% of OOD (resp., adversarial) samples are correctly rejected.

Benchmarks. Three medium- and large-sized datasets are considered: CIFAR-10, CIFAR-100 (Krizhevsky, 2009) and ImageNet-1K (using the ILSVRC2012 version). Three types of neural architectures are used to assess the generality of the SPADE approach: ResNet (He et al., 2016), ViT (Dosovitskiy et al., 2021), and VGG (Simonyan & Zisserman, 2015). The distance in latent space between a sample and its k-th nearest neighbor is calculated using the normalized L_2 distance, following Sun et al. (2022). OOD samples are sourced from near-OOD and far-OOD datasets, following (Cimpoi et al., 2014; Vaze et al., 2022; Horn et al., 2018; Wang et al., 2022; Bitterwolf et al., 2023), as detailed in the supplementary material (SM). Adversarial samples are generated by perturbing training samples using FGSM (Goodfellow et al., 2015) attacks, with a perturbation amplitude ε ranging from 0.001 to 0.004.

5 Experimental Results

OOD Detection (Q1). The performance of SPADE is illustrated in Table 1, focusing on the representative case of ImageNet-1K and considering a ResNet teacher. For the considered near-OOD datasets, the best method is MSP, whereas for far-OOD datasets, the best method is KNN. In all cases but one, SPADE-ResNet is slightly outperformed by KNN. In terms of rank (determined by the average of AUC and FPR95), SPADE ranks second best on both near- and far-OOD datasets.

[2]Note that the requirement for $(G^{(c,c')})^{-1}(1-\tau)$ to be sufficiently large is reminiscent of the clustering assumption that underpins semi-supervised learning (Rigollet, 2007).

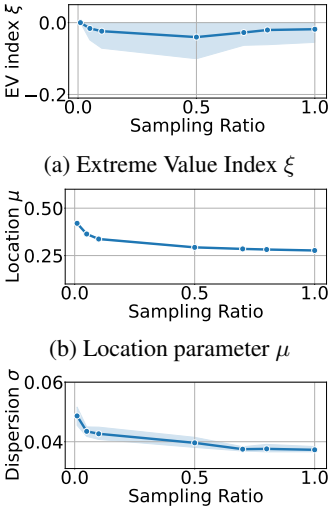
Table 1: OOD detection on ImageNet-1K: performance of SPADE (with ResNet teacher) compared to that of baselines MSP, ODIN, MDS and KNN in terms of AUC (the greater the better) and FPR95 (the lower the better; best performances in bold). The rank, averaged over far and near OOD datasets, is computed after the half sum of AUC and FPR95.

	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**

Table 2: OOD detection on CIFAR-10: performance of SPADE with teachers ResNet, VGG and ViT-B16, compared to that of baseline CIDER.

	TIN		MNIST		SVHN		Textures		Places-365	
	AUC ↑	FPR95 ↓	AUC ↑	FPR95 ↓	AUC ↑	FPR95 ↓	AUC ↑	FPR95 ↓	AUC ↑	FPR95 ↓
CIDER (d=512)	71.56	70.34	68.84	71.86	57.51	78.43	71.06	70.70	71.73	69.97
ResNet-18 (d=512)	90.41	32.14	93.646	21.91	92.17	22.94	91.97	25.68	91.03	30.41
VGG-16 (d=512)	76.04	66.84	89.37	35.12	81.53	44.06	80.47	50.73	72.14	77.27
ViT-B16 (d=384)	96.27	20.18	94.52	12.23	81.78	34.83	99.97	00.06	99.67	00.51

Sensitivity of OOD Detection (Q2). The impact of the considered teacher on the performance of the SPADE OOD test is illustrated in Table 2, comparing SPADE built on teachers ResNet, VGG, ViT-B16 with the CIDER baseline on CIFAR-10. These results show that the detection accuracy indeed depends on the teacher and its latent space, with SPADE-ViT-B16 dominating ResNet (except on SVHN) and ResNet strongly dominating VGG. Still, the performance does not merely depend on the size of the latent space. The discrepancy between the AUC and FPR95 indicators suggests that the optimal detection threshold varies depending on the dataset. Notably, all SPADE OOD detection tests outperform CIDER, which is based on KNN and specifically targets OOD detection. We shall come back to this remark in Section 6.



(a) Extreme Value Index ξ

(b) Location parameter μ

(c) Digression parameter σ

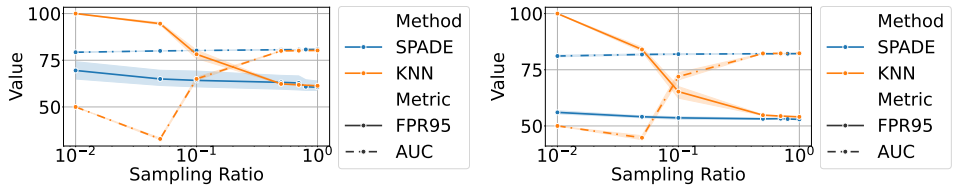
Figure 1: Stability of EVT parameter estimation wrt sampling ratio and estimation variance for one class of CIFAR-100 on ResNet-18.

Detection of Adversarial Samples (Q3). Table 3 reports the performance of SPADE (with a ResNet teacher) on the representative cases of CIFAR-10 and CIFAR-100, compared with MSP, MDS, KNN and CIDER. For all perturbation amplitudes, SPADE ranks first w.r.t. AUC (and second w.r.t. FPR95). In terms of FPR95, KNN ranks first on CIFAR-10, while MSP ranks first on CIFAR-100. Overall, SPADE behaves on par with, or better than, OOD detection methods w.r.t. the detection of adversarial examples. The slight AUC improvement suggests that SPADE may capture more subtle differences between in-distribution and adversarial samples. Conversely, the high FPR95 values suggest that SPADE tends to be overly cautious, rejecting true samples at the level of confidence where 95% adversarial samples are rejected.

Computational Frugality: Stability of GEV Models and SPADE Performances (Q4). The stability of the GEV models with respect to the fraction of training samples used to estimate the $\widehat{G}^{(c)}$ hyperparameters is illustrated in Fig. 1, using the ResNet teacher's latent space on CIFAR-100. The figure shows: i) a low sensitivity of the tail index (Fig. 1.(a)); ii) a decrease in μ with the sampling rate (Fig. 1.(b)); iii) a low sensitivity of the dispersion parameter σ (Fig. 1.(c)). Complementary

Table 3: Adversarial samples detection on CIFAR-10 and CIFAR-100, with perturbation amplitude from .001 to .004: comparison of SPADE (ResNet teacher) with baselines MSP, ODIN, MDS and KNN in terms of AUC (the greater the better) and FPR95 (the lower the better; best in bold).

		$\varepsilon = 0.001$		$\varepsilon = 0.002$		$\varepsilon = 0.003$		$\varepsilon = 0.004$		Average	
		AUC ↑	**FPR95** ↓	**AUC** ↑	**FPR95** ↓	**AUC** ↑	**FPR95** ↓	**AUC** ↑	**FPR95** ↓	**AUC** ↑	**FPR95** ↓
CIFAR-10	MSP (1)	81.68	79.08	81.90	78.70	82.06	79.14	82.20	78.10	81.96	78.76
	MDS (3)	81.46	69.34	81.51	68.76	81.57	69.04	81.61	69.42	81.54	69.14
	KNN (4)	85.65	**54.46**	85.75	**54.19**	85.85	**53.91**	85.92	54.29	85.79	**54.21**
	CIDER (5)	85.46	55.68	85.50	54.79	85.53	54.93	85.60	54.57	85.52	54.99
	SPADE	**85.96**	55.02	**86.06**	54.51	**86.15**	54.40	**86.24**	**53.78**	**86.10**	54.43
CIFAR-100	MSP (1)	83.24	**51.84**	83.39	**50.57**	83.54	**49.94**	83.68	**49.64**	83.46	**50.50**
	MDS (3)	60.34	82.93	60.29	83.01	60.25	82.60	60.19	83.01	60.27	82.89
	KNN (4)	83.54	56.83	83.67	56.39	83.79	55.64	83.89	55.13	83.72	56.00
	CIDER (5)	82.75	63.17	82.85	63.57	82.95	63.17	83.06	62.34	82.90	63.06
	SPADE	**84.33**	53.13	**84.45**	52.84	**84.56**	52.44	**84.66**	52.44	**84.50**	52.72



(a) near OOD (b) far OOD

Figure 2: Sensitivity analysis of the OOD detection on CIFAR-100 w.r.t. the subsampling rate of the training set: AUC (dashed line) and FPR95 (solid line) performances for SPADE (in blue) and KNN (Sun et al., 2022) (in orange; better seen in color).

results are provided in the SM, confirming that the lower bound on the adversarial perturbation amplitude is effectively positive.

As expected, the stability of the $\widehat{G}^{(c)}$ models results in stable OOD detection performances (AUC and FPR95) when subsampling the training set. Quite the contrary, the OOD detection performances of KNN are significantly deteriorated when aggressively subsampling the training set, particularly so on near-OOD (Fig. 2).

6 Position w.r.t. Related Work

The robustness lack of machine learning models with respect to adversarial and OOD examples is widely recognized as a major obstacle for ML applications in safety-critical domains (Salehi et al., 2022; Yang et al., 2024).

With respect to OOD examples, their detection takes inspiration from several areas of ML, ranging from learning with rejection (Bartlett & Wegkamp, 2008), to anomaly detection (Bulusu et al., 2020), novelty detection (Markou & Singh, 2003a;b), and open set recognition (Boult et al., 2019). To the best of our knowledge, the OOD detection problem was first formalized by Nguyen et al. (2015); the early MSP approach, observing the margin between the logits of the trained teacher and exploiting the fact that it behaves differently for in-distribution and OOD samples (Hendrycks & Gimpel, 2017), still is among the most effective ones (Table 1). Along the same line, ODIN exploits the gradient information to separate in- and out-of-distribution samples (Liang et al., 2018).
Some approaches address OOD detection as yet another supervised learning issue, treating OOD samples as belonging to an additional class and utilizing them to (re-)train the model (Du et al., 2022a; Zhang et al., 2022) (see also (Hsu et al., 2020)).

Quite a few other methods are based on the so-called manifold assumption; the challenge, then, is to identify the representation that best characterizes the manifold on which real samples lie. One option is to consider the latent representation of an auto-encoder (AE) trained solely on real samples. As empirically shown by e.g. Ren et al. (2019); Liu et al. (2021); Du et al. (2022a), the reconstruction

error of OOD samples through the AE is much higher than for real samples; an effective OOD test can thus be based on this error. On the positive side, this reconstruction error defines a general criterion and does not depend on the teacher; on the negative side, it does not leverage class information. Another option is based on modeling the behavior of true samples in the penultimate layer of the neural net, using probabilistic models (Du et al., 2022b; Liu et al., 2021; Ren et al., 2019).

Finally, the option most closely related to SPADE is to consider the latent representation of the teacher itself, as is done in distance-based OOD detection approaches (Papernot & McDaniel, 2018; Lee et al., 2018), particularly in KNN (Sun et al., 2022). The difference is that KNN exploits the distance z of a sample to its k-th nearest neighbor in the training set, while SPADE exploits $\widehat{G}^{(c)}(z_c)$. A tentative interpretation for the better performance of KNN (Table 1) is the superior bias-variance trade-off in the empirical test based on z compared to the test based on $\widehat{G}^{(c)}(z_c)$. While learning a parametric GEV model achieves some regularization, this model is only trained from examples within the class c, meaning it operates with one or two orders of magnitude less data.

The problem of dealing with adversarial examples differs from that of detecting OOD examples, as adversarial examples are deliberately crafted to deceive the teacher (Szegedy et al., 2014; Goodfellow et al., 2015; Madry et al., 2018; Croce & Hein, 2020). Knowing their structure allows for the design of specific defense strategies, such as adversarial training (Goodfellow et al., 2015; Madry et al., 2018), which incorporates adversarial examples into the training process (Yan et al., 2018; Zhang et al., 2019; Wang et al., 2020). Other notable defense strategies include adversarial architectures (Hosseini et al., 2020; Huang et al., 2023), adversarial regularization (Mao et al., 2019; Liu et al., 2023), and data augmentation methods (Carmon et al., 2019; Wang et al., 2023).

In contrast, SPADE neither requires additional information nor retrains the classifier to defend against adversarial examples. It employs the same agnostic strategy for both adversarial and OOD threats: it characterizes the true samples and abstains from making a decision if the example in question appears extreme compared to the true (ID) samples, based on the chosen confidence level.

7 Conclusion and Perspectives

The main contribution of the paper, SPADE (*Sample-efficient ProbAbilistic Detection using Extreme value theory*), is a formal test designed to detect examples appearing to be extreme w.r.t. training samples, enabling the classifier to abstain from making decisions on such extreme examples. This test's ability to accurately detect OOD and adversarial samples has been empirically demonstrated, with similar performances as the prominent state of the art approaches. Furthermore, the computational complexity and the stability of the proposed detection test have been empirically established.

The proposed test, similar to distance-based OOD detection approaches, exploits the latent distance between the given example and its nearest neighbor in the training set. Its originality lies in leveraging Extreme Value Theory (Fisher & Tippett, 1928) to provide a formal characterization of the training samples. This characterization offers two key benefits: first, it yields provable guarantees for detecting adversarial examples, subject to the adversarial perturbation amplitude to be lower bounded; second, it provides some new hints into the key aspects governing the teacher robustness.[3]

This approach opens several perspectives, related with making classifiers more robust and better understanding the key robustness factors. A short term perspective is to extend the generalized extreme value test to some of the empirical criteria used in the OOD detection literature; one such criterion is the score margin involved in MSP (Hendrycks & Gimpel, 2017). Another perspective is to enhance the classifier training loss to favor the robustness of the latent space, e.g. to consider the optimization of the Lipschitz constant of the classifier embedding besides its variation and informativeness as done in CIDER (Ming et al., 2023).

Our long-term goal is to investigate whether *safe* example behaviors can be identified in the latent space and whether these behaviors can be certified, as a step toward the certification of neural networks.

[3]For instance, CIDER Ming et al. (2023) involves the optimization of the variation and informativeness of the teacher latent space. The SPADE analysis suggests that besides these two factors, the regularity of the teacher (its Lipschitz constant) also matters. A possible interpretation for why SPADE outperforms CIDER is that the optimization of the variation and informativeness might adversely affect this Lipschitz constant.

REFERENCES

A. A. Balkema and L. de Haan. Residual Life Time at Great Age. *The Annals of Probability*, 1974.

Peter L. Bartlett and Marten H. Wegkamp. Classification with a reject option using a hinge loss. *JMLR*, 2008.

Julian Bitterwolf, Maximilian Mueller, and Matthias Hein. In or out? fixing imagenet out-of-distribution detection evaluation. *ICML*, 2023.

T. E. Boult, S. Cruz, A.R. Dhamija, M. Gunther, J. Henrydoss, and W.J. Scheirer. Learning and the unknown: Surveying steps toward open world recognition. *AAAI*, 2019.

Saikiran Bulusu, Bhavya Kailkhura, Bo Li, Pramod K. Varshney, and Dawn Song. Anomalous example detection in deep learning: A survey. *IEEE Access*, 2020.

Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. *NeurIPS*, 2019.

M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. *CVPR*, 2014.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. *ICML*, 2020.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021.

Xuefeng Du, Xin Wang, Gabriel Gozum, and Yixuan Li. Unknown-aware object detection: Learning what you don't know from videos in the wild. *CVPR*, 2022a.

Xuefeng Du, Zhaoning Wang, Mu Cai, and Yixuan Li. Vos: Learning what you don't know by virtual outlier synthesis. *ICLR*, 2022b.

Adam Dziedzic, Stephan Rabanser, Mohammad Yaghini, Armin Ale, Murat A. Erdogdu, and Nicolas Papernot. p-dknn: Out-of-distribution detection through statistical testing of deep representations. *ArXiv*, 2022.

Sebastian Farquhar and Yarin Gal. What 'out-of-distribution' is and is not. *NeurIPS ML Safety Workshop*, 2022.

Ronald Aylmer Fisher and L. H. C. Tippett. Limiting forms of the frequency distribution of the largest or smallest member of a sample. *Mathematical Proceedings of the Cambridge Philosophical Society*, 1928.

Joshua P. French, Piotr Kokoszka, Stilian A. Stoev, and Lauren Hall. Quantifying the risk of heat waves using extreme value theory and spatio-temporal functional data. *Comput. Stat. Data Anal.*, 2019.

Boris Gnedenko. Sur la distribution limite du terme maximum d'une serie aleatoire. *Annals of mathematics*, 1943.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *ICLR*, 2015.

Laurens Haan and Ana Ferreira. *Extreme Value Theory: An Introduction*. Springer Series in Operations Research and Financial Engineering. Springer, New York, 2006.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CVPR*, 2016.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *ICLR*, 2017. Alias (1) used in Table 1.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. *CVPR*, 2018.

Ramtin Hosseini, Xingyi Yang, and Pengtao Xie. Dsrna: Differentiable search of robust neural architectures. *CVPR*, 2020.

Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. *CVPR*, 2020.

Shihua Huang, Zhichao Lu, Kalyanmoy Deb, and Vishnu Naresh Boddeti. Revisiting residual networks for adversarial robustness. *CVPR*, 2023.

Badr Youbi Idrissi, Diane Bouchacourt, Randall Balestriero, Ivan Evtimov, Caner Hazirbas, Nicolas Ballas, Pascal Vincent, Michal Drozdzal, David Lopez-Paz, and Mark Ibrahim. Imagenet-x: Understanding model mistakes with factor of variation annotations. *ICLR*, 2023.

Wenyu Jiang, Yuxin Ge, Hao Cheng, Mingcai Chen, Shuai Feng, and Chongjun Wang. Read: Aggregating reconstruction error into out-of-distribution detection. *AAAI*, 2023.

Alex Krizhevsky. Learning multiple layers of features from tiny images. *Technical Report*, 2009.

Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *NeurIPS*, 2018. Alias (3) used in Table 1.

Jingyao Li, Pengguang Chen, Shaozuo Yu, Zexin He, Shu Liu, and Jiaya Jia. Rethinking out-of-distribution (ood) detection: Masked image modeling is all you need, 2023.

Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *ICLR*, 2018. Alias (2) used in Table 1.

Hong Liu, Zhun Zhong, Nicu Sebe, and Shin'ichi Satoh. Mitigating robust overfitting via self-residual-calibration regularization. *Artificial Intelligence*, 2023.

Weitang Liu, Xiaoyun Wang, John D. Owens, and Yixuan Li. Energy-based out-of-distribution detection. *NeurIPS*, 2021.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *ICLR*, 2018.

Chengzhi Mao, Ziyuan Zhong, Junfeng Yang, Carl Vondrick, and Baishakhi Ray. Metric learning for adversarial robustness. *NeurIPS*, 2019.

Markos Markou and Sameer Singh. Novelty detection: a review—part 1: statistical approaches. *Signal Processing*, 2003a.

Markos Markou and Sameer Singh. Novelty detection: a review—part 2: neural network based approaches. *Signal Processing*, 2003b.

Yifei Ming, Yiyou Sun, Ousmane Dia, and Yixuan Li. How to exploit hyperspherical embeddings for out-of-distribution detection? *ICLR*, 2023. Alias (5) used in Table 1.

Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. *CVPR*, 2015.

Nicolas Papernot and Patrick D. McDaniel. Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. *ArXiv*, 2018.

J. Pickands. Statistical Inference Using Extreme Order Statistics. *The Annals of Statistics*, 1975.

Jie Ren, Peter J. Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. *NeurIPS*, 2019.

Philippe Rigollet. Generalization error bounds in semi-supervised classification under the cluster assumption. *JMLR*, 2007.

Mohammadreza Salehi, Hossein Mirzaei, Dan Hendrycks, Yixuan Li, Mohammad Hossein Rohban, and Mohammad Sabokrou. A unified survey on anomaly, novelty, open-set, and out-of-distribution detection: Solutions and future challenges. *TMLR*, 2022.

Alban Siffer, Pierre-Alain Fouque, Alexandre Termier, and Christine Largouet. Anomaly detection in streams with extreme value theory. *KDD*, 2017.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *ACPR*, 2015.

Mark Smith, Steven Reece, Stephen J. Roberts, and Iead Rezek. Online maritime abnormality detection using gaussian processes and extreme value theory. *ICDM*, 2012.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. *ICML*, 2022. Alias (4) used in Table 1.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *ICLR*, 2014.

Sagar Vaze, Kai Han, Andrea Vedaldi, and Andrew Zisserman. Open-set recognition: a good closed-set classifier is all you need? *ICLR*, 2022.

Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. *CVPR*, 2022.

Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. *ICLR*, 2020.

Zekai Wang, Tianyu Pang, Chao Du, Min Lin, Weiwei Liu, and Shuicheng Yan. Better diffusion models further improve adversarial training. *ICML*, 2023.

Ziang Yan, Yiwen Guo, and Changshui Zhang. Deep defense: Training dnns with improved adversarial robustness. *NeurIPS*, 2018.

Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Dan Hendrycks, Yixuan Li, and Ziwei Liu. Openood: Benchmarking generalized out-of-distribution detection. *NeurIPS*, 2022.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *IJCV*, 2024.

Haotian Ye, Chuanlong Xie, Tianle Cai, Ruichen Li, Zhenguo Li, and Liwei Wang. Towards a theoretical framework of out-of-distribution generalization. *NeurIPS*, 2021.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. *ICML*, 2019.

Jingyang Zhang, Nathan Inkawhich, Randolph Linderman, Yiran Chen, and Hai Li. Mixture outlier exposure: Towards out-of-distribution detection in fine-grained environments. *WACV*, 2022.

Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Yixuan Li, Ziwei Liu, Yiran Chen, and Hai Li. Openood v1.5: Enhanced benchmark for out-of-distribution detection. *arXiv*, 2023.