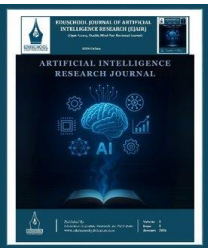
EDUSCHOOL JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH (EJAIR)
(Open Access, Double-Blind Peer Reviewed Journal)
ISSN Online: 3139-7336



Adversarial Robustness Under Distribution Shift: A Unified View of Detection and Defense

Tintu George

Assistant Professor, Department of BCA AI, Sri Ramakrishna College of Arts & Science, Coimbatore, Tamil Nadu, India

Article information

Received: 24th July 2026
Received in revised form: 27th August 2026
Accepted: 30th September 2026
Available online: 17th November 2026

Volume: 1
Issue: 3
DOI: https://doi.org/10.63090/EJAIR/3139.7336.0008

Abstract

Adversarial robustness and out-of-distribution detection have developed as separate research programmes, with separate benchmarks and metrics, although both concern how a classifier behaves on inputs that its training data do not represent. This article reviews them together with work on natural distribution shift and argues that all three study points in one space of test conditions. The space has two axes. The first separates covariate shift, where the label set is unchanged and the correct response is the correct label, from semantic shift, where the input belongs to no training class and the correct response is abstention. The second orders conditions by how strongly the shifted inputs are selected against the model, from random corruption and dataset re-collection, through natural images filtered for model failure, to white-box perturbations chosen by an adaptive adversary. Attack and evaluation methods, natural shift benchmarks, post hoc and training-based detectors, the robustness-accuracy trade-off and certified defenses are reviewed and placed in this space. A relation from the Gaussian noise analysis of adversarial examples is used to connect the random and adversarial ends of the selection axis. An evaluation protocol is then proposed that measures a classifier and its detector as one system in every cell, including attacks written against the detector. The framework and protocol are proposed and have not yet been tested.

Keywords:- Adversarial Robustness, Distribution Shift, Out-of-Distribution Detection, Threat Model, Randomized Smoothing, Evaluation Protocol

.

I. INTRODUCTION

A deep network trained on a fixed distribution is evaluated, in most published work, on a held-out sample of that same distribution. Deployment rarely respects this assumption. Images arrive blurred, compressed or taken under unfamiliar lighting, objects appear in renditions never seen during training, and some inputs belong to no class the model was asked to recognise. A further kind of input is constructed on purpose, since small perturbations found by an optimiser can change the prediction of an accurate image classifier while leaving the image visually unchanged [1], [2]. Each situation has produced its own literature.

The three literatures use different measures. Adversarial robustness is measured as accuracy under a worst-case perturbation inside an $\ell\infty$ or $\ell2$ ball. Most defenses are built on the min-max formulation of adversarial training [3]. Robustness to natural shift is measured on corrupted or re-collected test sets such as ImageNet-C [4] and ImageNetV2 [5]. Out-of-distribution (OOD) detection is measured by how well a confidence score separates in-distribution test images from images of unseen classes, starting from the maximum softmax probability baseline

[6]. The three communities report different metrics, and a model presented as strong under one of them is seldom evaluated under the other two.

There are good reasons to treat them together. One early detector, based on class-conditional Gaussian models of network features, was shown to flag both unseen-class inputs and adversarial examples with the same score [7]. An analysis of Gaussian noise has tied the error rate of a classifier under random corruption to the distance from a clean input to its nearest misclassified point [8]. A detector is itself a function of the input. As a result, an attacker who knows it can perturb an OOD input until the detector accepts it. A deployed system therefore faces natural shift, semantic novelty and deliberate manipulation at the same time, and an evaluation that measures them one at a time cannot see the interactions between them.

This article reviews the three literatures and proposes a unified view of them, beginning in Section II with adversarial attacks and with the evaluation failures that shaped current practice in that field. Section III covers natural distribution shift, and Section IV covers OOD detection and its exposure to attack. Section V proposes a threat model with two axes, the type of shift and the strength of selection against the model. Section VI discusses defenses, the robustness-accuracy trade-off and certification. Section VII proposes an evaluation protocol and states the limitations of the proposal.

II. ADVERSARIAL EXAMPLES AND THE EVALUATION PROBLEM

Szegedy et al. reported in 2014 that imperceptible perturbations could cause misclassification [1]. They also found that the same perturbed images often fooled networks trained separately [1]. Goodfellow et al. attributed the effect to the nearly linear behaviour of networks in high dimension [2], and their fast gradient sign method (FGSM) moves the input one step of size ε along the sign of the loss gradient. Madry et al. recast attack and defense as one saddle point problem [3], in which the objective is the minimum over parameters θ of the expected maximum over $\|\delta\| \le \varepsilon$ of the loss $\mathcal{L}(f_\theta(x+\delta), y)$. Projected gradient descent (PGD) solves the inner maximisation. Carlini and Wagner wrote attacks as optimisation problems under $\ell 0$, $\ell 2$ and $\ell\infty$ distances and used them to defeat defensive distillation, which had resisted the earlier attacks [9].

Much of the later history concerns evaluation. Athalye et al. examined the non-certified white-box defenses accepted at ICLR 2018 and found that seven of the nine relied on obfuscated gradients, a form of gradient masking that defeats gradient-based attacks without removing the adversarial examples [10]. With attacks adapted to each mechanism they circumvented six of the seven completely and one partially. Tramer et al. later broke thirteen published defenses, each with an attack written for that defense, and set out the methodology an adaptive attack requires, arguing that each attack has to be tailored to the defense rather than run as a fixed procedure [11]. A defense therefore has to be evaluated by an attacker who knows it. An added OOD detector enlarges the surface that such an attacker can target.

Standardised attacks removed the worst cases of weak evaluation. AutoAttack has four components [12]. Two are parameter-free variants of PGD with an automatic step size schedule, and the other two are the minimum-norm FAB attack and the query-based Square attack. Applied to more than fifty published models, it gave a lower robust accuracy than the original papers in all but one case, often by more than ten percentage points [12]. RobustBench maintains leaderboards built on AutoAttack for $\ell\infty$ and $\ell 2$ threat models and for common corruptions [13]. They cover CIFAR-10, CIFAR-100 and ImageNet. Entries are restricted to models for which the attack is known to be reliable, which excludes randomised defenses and defenses that run an optimisation loop at test time. Almost all threat models in this literature are ℓp balls around a clean input, chosen for tractability. Real adversaries are not confined to such balls.

III. NATURAL DISTRIBUTION SHIFT

Most natural shift benchmarks for vision are built on ImageNet [14] and keep its label set or a subset of it, so that the same classifier can be tested without retraining. ImageNet-C applies fifteen corruption types, grouped into noise, blur, weather and digital artefacts, at five severity levels each [4]. Each validation image therefore has 75 corrupted versions. The benchmark reports corruption error normalised by that of AlexNet. ImageNetV2 repeats the original collection procedure on new images and finds accuracy drops of 11% to 14% for ImageNet models, which the authors attribute to the new images being slightly harder rather than to adaptive overfitting on the original test set [5].

Other benchmarks change the kind of image. ImageNet-R contains 30,000 renditions of 200 ImageNet classes, including paintings, cartoons, sculptures and embroidery [15]. ImageNet-A contains real, unmodified photographs gathered by adversarial filtration, which keeps the web images that a fixed classifier gets wrong, and its companion ImageNet-O applies the same filtering to images of classes outside ImageNet to build an OOD detection set [16]. On ImageNet-A a DenseNet-121 reaches about 2% accuracy, and its OOD detection on

ImageNet-O is near chance. WILDS collects ten datasets with shifts that occur in practice, among them hospitals, camera trap locations and satellite regions, and defines its splits by those domains [17].

Two empirical regularities organise this literature. Taori et al. evaluated 204 ImageNet models under 213 test conditions [18]. They found little or no transfer of robustness from synthetic shifts to natural ones, with training on larger and more diverse data as the main exception [18]. Miller et al. found that, across many models and shifts, accuracy on the shifted distribution is close to a linear function of in-distribution accuracy after a probit transform of both [19]. A method improves robustness only if it lifts shifted accuracy above the line predicted from its in-distribution accuracy, a quantity Taori et al. call effective robustness [18]. Hendrycks et al. reached a less negative conclusion on their own datasets [15]. They found that larger models and some artificial augmentations improve robustness to real shifts but that no evaluated method helps consistently [15]. The usual explanation for poor transfer is shortcut learning, in which a network adopts decision rules, such as reliance on texture or background, that suffice on the benchmark and fail outside it [20].

IV. OUT-OF-DISTRIBUTION DETECTION AND ITS EXPOSURE TO ATTACK

OOD detection asks for a score s(x) that is high for in-distribution inputs and low for inputs whose class lies outside the training label set. The maximum softmax probability (MSP) of a trained classifier gives the standard baseline [6]. It needs no retraining. ODIN sharpens the separation with temperature scaling of the softmax and a small input perturbation that increases the maximum softmax score, a step with the same form as FGSM taken in the opposite direction [21]. The Mahalanobis detector fits class-conditional Gaussians with a shared covariance to intermediate features and scores an input by its distance to the closest class mean [7]. The energy score replaces the softmax maximum with :

$$E(x) = -T \log \sum_i \exp\left(\frac{f_i(x)}{T}\right) \quad (1)$$

computed from the logits $f_i(x)$ at temperature T [22].

Later post hoc methods clip penultimate activations before computing the score [23], or use the distance from an input to its k-th nearest training feature in a normalised embedding [24]. Outlier exposure takes a different route and trains the classifier to give uniform predictions on an auxiliary dataset of outliers [25]. Table 1 lists these families.

Table 1. Families of out-of-distribution detector

Family	Score computed from	Retraining needed	Auxiliary outliers	Representative work
Softmax confidence	Maximum softmax probability	No	No	[6]
Temperature and input perturbation	Scaled softmax after one gradient step	No	For tuning	[21]
Feature density	Mahalanobis distance to class means	No	For tuning	[7]
Energy	Log-sum-exp of the logits	Optional	Optional	[22]
Activation shaping	Clipped penultimate activations	No	No	[23]
Nearest neighbour	k-th neighbour distance in feature space	No	No	[24]
Outlier exposure	Confidence after training on outliers	Yes	Yes	[25]
Adversarial confidence training	Confidence on perturbed outliers	Yes	Synthetic or real	[26], [27]

OpenOOD gathers detectors and datasets under one protocol [28]. It separates near-OOD sets, whose classes resemble the training classes, from far-OOD sets [28]. Yang et al. place OOD detection beside anomaly detection, novelty detection, open set recognition and outlier detection in a single framework of generalised OOD detection [29]. Large pretrained transformers raise performance on hard near-OOD pairs such as CIFAR-100 against CIFAR-10 [30]. Apart from the last row, none of the scores in Table 1 was designed with an attacker in mind, and each is computed from a network that an attacker can differentiate through.

In 2019 Hein et al. proved that ReLU networks produce arbitrarily high-confidence predictions on inputs far from the training data, because scaling an input by a large factor drives the logits along a fixed linear direction [26]. Carlini and Wagner bypassed ten detectors of adversarial examples by adding each detector to the attack objective [31]. The same construction applies to OOD detectors. A perturbation that raises s(x) on an OOD input turns it into an accepted input, and a perturbation that lowers s(x) on an in-distribution input causes a false alarm. Provable guarantees exist for restricted settings, such as low confidence far from the training data under a density model [32] and low confidence within an $\ell\infty$ ball around OOD inputs using interval bound propagation [27]. These guarantees hold at small radii and have been shown on small datasets such as CIFAR-10.

V. A UNIFIED THREAT MODEL

The framework proposed here describes any test condition by two coordinates. The first, following the distinction drawn in [29], is the type of shift. Covariate shift changes the input distribution while keeping the label set Y fixed, so the correct response is the correct label. Semantic shift supplies inputs whose true class lies outside Y, so the correct response is abstention. The second coordinate is the strength of selection against the model, which runs from random sampling of shifted inputs, through filtering of natural inputs by model failure, to white-box optimisation by an adversary who knows both the classifier and the detector. Fig. 1 shows the resulting grid with existing benchmarks placed in its cells.

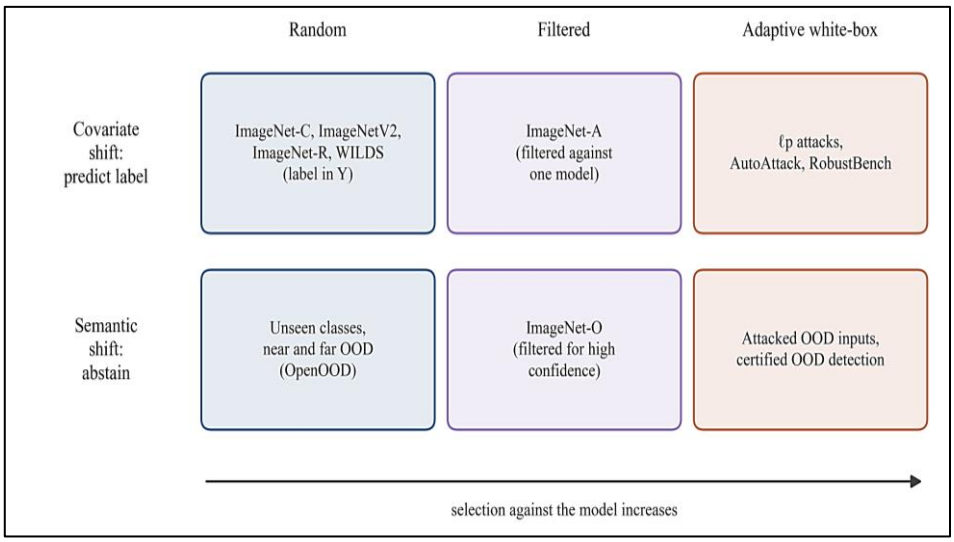


Fig. 1: Schematic of the proposed two-axis threat model. Rows give the type of shift and columns give the strength of selection against the model, with representative benchmarks in each cell: corruptions, renditions and re-collected sets [4], [15], [5], [17], filtered sets [16], norm-bounded attacks [12], [13], unseen-class benchmarks [28] and attacked OOD inputs [27].

Formally, a test condition is a triple (Q, Δ, A), where Q is a source distribution, Δ(x) is the set of admissible modifications of an input x, and A is the rule that selects one element of Δ(x). The system under test is a pair (f, s) with a threshold τ. It returns f(x) when $s(x) \geq \tau$ and abstains otherwise. Its risk is the expected loss of the selected input, with x drawn from Q, where the loss has separate costs for a wrong label on an accepted covariate input, an accepted semantic input and a rejected covariate input, including a perturbed one.

Standard adversarial robustness is the case Q = P, the training distribution, with A the worst case over an ℓp ball. Natural shift robustness takes A as the identity and Q as a corrupted or re-collected distribution. OOD detection also keeps A as the identity but takes Q supported outside Y. Table 2 records these cases and the intermediate ones.

Table 2. Placement of test conditions in the proposed two-axis threat model

Condition	Shift type	Selection	Admissible set Δ(x)	Correct response	Examples
In-distribution test	None	Random	The input itself	Label	[14]
Common corruptions	Covariate	Random	15 corruptions, 5 severities	Label	[4]
Re-collection and renditions	Covariate	Random	New sample from Q	Label	[5], [15], [17]
Filtered natural images	Covariate	Against one fixed model	Unmodified natural images	Label	[16]
Norm-bounded attack	Covariate	Adaptive white-box	ℓp ball of radius ε	Label	[3], [12]
Attack on the detector	Covariate	Adaptive white-box	ℓp ball, score as target	Label, no abstention	[31]
Unseen classes	Semantic	Random	New sample outside Y	Abstain	[6], [28]
Filtered OOD images	Semantic	Against one fixed model	Unmodified natural images	Abstain	[16]
Attacked OOD inputs	Semantic	Adaptive white-box	ℓp ball around an OOD input	Abstain	[26], [27]

The two ends of the selection axis are not independent. Gilmer et al. considered isotropic Gaussian noise with standard deviation σ per coordinate and the error rate μ of a classifier on noisy copies of a clean input x [8]. If the set of misclassified points is a half space, the ℓ2 distance from x to its nearest error equals:

$$\varepsilon^* = -\sigma\Phi^{-1}(\mu) \tag{2}$$

where Φ is the standard normal distribution function. Through the Gaussian isoperimetric inequality the half space is the most favourable shape, so error sets of other shapes with the same measure lie, in a median sense, no further away from the noisy copies of x [8]. Fig. 2 plots ε* for three noise levels. At σ = 0.1 an error rate of 1% under noise already places an error within an ℓ2 distance of about 0.23. At μ = 10% the distance falls to about 0.13.

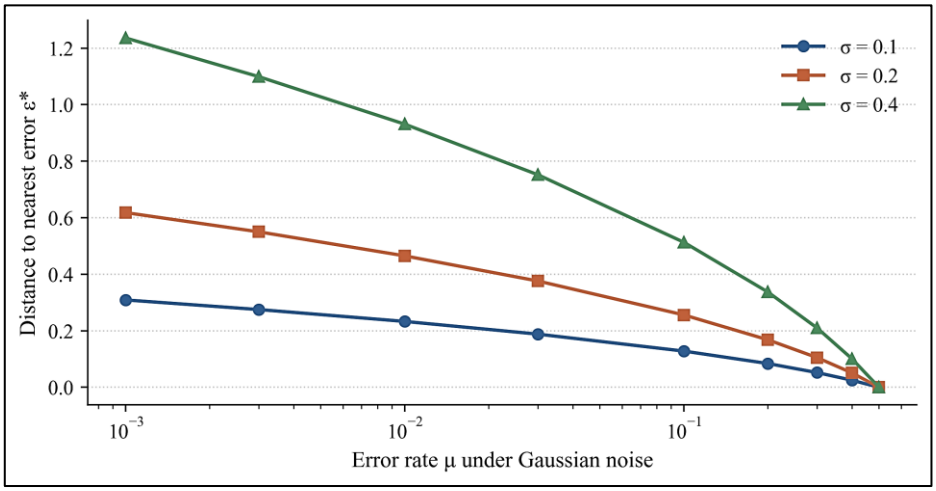


Fig. 2: Distance from a clean input to its nearest error, $\varepsilon^* = -\sigma\Phi^{-1}(\mu)$, as a function of the error rate μ under isotropic Gaussian noise with standard deviation σ, computed from the half space relation of [8]. The horizontal axis is logarithmic.

A model with measurable error under random noise is therefore expected, under the half-space approximation, to have adversarial errors close by, and Gilmer et al. accordingly treat adversarial and corruption robustness as two views of one phenomenon [8]. ImageNet-A and ImageNet-O were built by keeping natural images on which a fixed model failed [16]. The adversary here selects inputs without modifying them and uses no gradients. It therefore sits between random shift and white-box attack. The rows of Table 2 for attacks on the detector and for attacked OOD inputs are the cells that current practice leaves sparsely populated.

VI. DEFENSES, TRADE-OFFS AND CERTIFICATION

Adversarial training with PGD and its variants underlies most of the leading entries on the RobustBench leaderboards [3], [13]. It costs clean accuracy. Tsipras et al. constructed a simple distribution on which any classifier with high standard accuracy must have low robust accuracy [33]. TRADES writes the robust error as

the sum of the natural error and a boundary error, and trains with a cross-entropy term plus a Kullback-Leibler term between the predictions on x and on its worst-case perturbation, weighted by a parameter that sets the balance [34].

In the two-axis view the trade-off has a second dimension. Adversarial training constrains covariate inputs within an ℓp ball. It places no constraint on semantic inputs, and a model trained only to keep its label inside the ball has no reason to lower its confidence on an unseen class. Outlier exposure trains low confidence on semantic inputs without any constraint on perturbations [25]. Hein et al. combined the two by enforcing low confidence on perturbed noise inputs [26], and Bitterwolf et al. made this certifiable within ℓ∞ balls around outliers [27]. No published defense known to the authors covers all six cells of Fig. 1. The cost of defending one cell in terms of another has been reported mainly for single pairs of cells, most often clean accuracy against ℓ∞ robustness.

Certification replaces an empirical attack with a guarantee. Randomized smoothing classifies a point by the most probable output of a base classifier under Gaussian noise with standard deviation σ. Cohen et al. proved that the smoothed classifier is constant within the ℓ2 radius:

$$R = \frac{\sigma}{2}[\Phi^{-1}(p_A) - \Phi^{-1}(p_B)] \quad (3)$$

where p_A is a lower bound on the probability of the top class and p_B an upper bound on that of the runner-up [35]. The bound is tight for Gaussian noise. With $p_B = 1 - p_A$ the radius becomes $\sigma\,\Phi^{-1}(p_A)$. Fig. 3 plots this radius for the three noise levels Cohen et al. used on ImageNet.

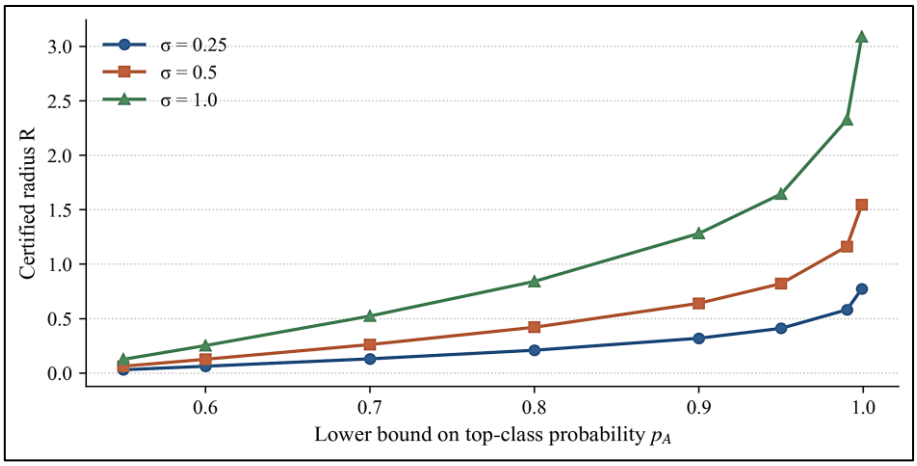


Fig. 3: Certified ℓ2 radius of a Gaussian smoothed classifier, $R = \sigma\Phi^{-1}(p_A)$, for the case $p_B = 1 - p_A$, computed from the bound of [35] for the three noise levels used there on ImageNet.

In Fig. 3 the radius exceeds σ only when p_A passes $\Phi(1)$, about 0.84. It grows slowly after that, reaching 2.33σ at $p_A = 0.99$. A larger σ buys a larger radius but makes the base classifier work on noisier inputs, and on ImageNet Cohen et al. reported a certified top-1 accuracy of 49% at an ℓ2 radius of 0.5 [35]. Smoothing turns accuracy under Gaussian noise, a random covariate condition, into a guarantee against white-box ℓ2 attack. This is a converse of the relation of Fig. 2, applied to the smoothed classifier. Pairing a smoothed classifier with a smoothed detector score has received less study than the smoothed classifier alone. Certified radii of this size are also small compared with the change that a rendition or a weather corruption makes to an image.

VII. PROPOSED EVALUATION PROTOCOL AND LIMITATIONS

The protocol proposed here reports a classifier and its detector together, as one system (f, s, τ), in every cell of Fig. 1. Six measurements are required. Table 2 lists the cells they cover. Clean accuracy on the in-distribution test set fixes the reference point. Accuracy under random covariate shift is reported per corruption and severity and on re-collected sets, together with its distance above or below the linear trend of shifted against clean accuracy, so that gains are not credited to higher clean accuracy alone [18], [19]. Accuracy on filtered sets is reported with the identity of the filtering model. Robust accuracy under AutoAttack is reported at several radii rather than one [12]. Detection of semantic shift is reported separately for near-OOD and far-OOD sets under a shared protocol such as OpenOOD [28]. The sixth measurement attacks the detector, perturbing in-distribution inputs to lower s(x) and OOD inputs to raise it, with the attack written against the whole system, including any input preprocessing such as the ODIN step [21], [11].

A single summary needs a fixed threshold. The protocol sets τ on clean in-distribution validation data at a stated true positive rate, such as 95%. It is then held fixed across all cells. For each cell it then reports the

acceptance rate and the selective risk, which is the error rate among accepted inputs. Where a certificate exists, the certified fraction at each radius is reported beside the empirical figure, following the practice of Cohen et al. for accuracy [35].

Five limitations apply. The framework has not been applied to any model, so the expectation that it changes the ranking of methods relative to single-axis evaluation is untested. The selection axis is ordinal rather than metric, and two filtered datasets built against different models cannot be compared in strength. Semantic shift is defined relative to a label set, while many deployed inputs are partly novel, for example a known object in an unfamiliar context. The binary first axis does not describe such inputs. The attacks in the sixth measurement have to be adapted to each detector, which makes the protocol expensive, and like every empirical robustness figure its results give only an upper bound on true performance [11]. Almost all the evidence reviewed here comes from image classification, and its extension to language models, where shift and attack act on text rather than pixels, is open.

VIII. CONCLUSION

Adversarial robustness, robustness to natural shift and OOD detection can be read as studying the same object, the behaviour of a model away from its training distribution. They differ in the type of shift and in how strongly the shifted inputs are selected against the model. Most defenses cover one or two cells of the grid in Fig. 1, detectors are usually reported without the adaptive attacks that adversarial robustness research learned to require, and, under a half-space approximation, the relation plotted in Fig. 2 ties the random and adversarial ends of the selection axis together.

The proposed protocol measures one system in every cell, with a fixed threshold and with attacks written against the detector. Running it on the public models collected by RobustBench and OpenOOD is the next step [13], [28]. Safety methods that act on model behaviour, such as alignment through self-critique and revision as discussed in this journal [36], assume that a model behaves on deployment inputs as it did during development. The measurements proposed here give one way to check that assumption.

REFERENCES

[1] C. Szegedy et al., “Intriguing properties of neural networks,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2014.

[2] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015.

[3] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018.

[4] D. Hendrycks and T. Dietterich, “Benchmarking neural network robustness to common corruptions and perturbations,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019.

[5] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, “Do ImageNet classifiers generalize to ImageNet?,” in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, PMLR, vol. 97, 2019, pp. 5389–5400.

[6] D. Hendrycks and K. Gimpel, “A baseline for detecting misclassified and out-of-distribution examples in neural networks,” in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017.

[7] K. Lee, K. Lee, H. Lee, and J. Shin, “A simple unified framework for detecting out-of-distribution samples and adversarial attacks,” in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 31, 2018.

[8] J. Gilmer, N. Ford, N. Carlini, and E. Cubuk, “Adversarial examples are a natural consequence of test error in noise,” in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, PMLR, vol. 97, 2019, pp. 2280–2289.

[9] N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” in *Proc. IEEE Symp. Secur. Privacy (SP)*, 2017, pp. 39–57.

[10] A. Athalye, N. Carlini, and D. Wagner, “Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples,” in *Proc. 35th Int. Conf. Mach. Learn. (ICML)*, PMLR, vol. 80, 2018, pp. 274–283.

[11] F. Tramer, N. Carlini, W. Brendel, and A. Madry, “On adaptive attacks to adversarial example defenses,” in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 1633–1645.

[12] F. Croce and M. Hein, “Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks,” in *Proc. 37th Int. Conf. Mach. Learn. (ICML)*, PMLR, vol. 119, 2020, pp. 2206–2216.

[13] F. Croce et al., “RobustBench: A standardized adversarial robustness benchmark,” in *Proc. Neural Inf. Process. Syst. Track Datasets Benchmarks*, vol. 1, 2021.

[14] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, “ImageNet: A large-scale hierarchical image database,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2009, pp. 248–255.

[15] D. Hendrycks et al., “The many faces of robustness: A critical analysis of out-of-distribution generalization,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 8320–8329.

[16] D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, and D. Song, “Natural adversarial examples,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 15257–15266.

[17] P. W. Koh et al., “WILDS: A benchmark of in-the-wild distribution shifts,” in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, PMLR, vol. 139, 2021, pp. 5637–5664.

[18] R. Taori, A. Dave, V. Shankar, N. Carlini, B. Recht, and L. Schmidt, “Measuring robustness to natural distribution shifts in image classification,” in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 18583–18599.

[19] J. P. Miller et al., "Accuracy on the line: On the strong correlation between out-of-distribution and in-distribution generalization," in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, PMLR, vol. 139, 2021, pp. 7721–7735.
[20] R. Geirhos et al., "Shortcut learning in deep neural networks," *Nat. Mach. Intell.*, vol. 2, no. 11, pp. 665–673, 2020.
[21] S. Liang, Y. Li, and R. Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018.
[22] W. Liu, X. Wang, J. D. Owens, and Y. Li, "Energy-based out-of-distribution detection," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 21464–21475.
[23] Y. Sun, C. Guo, and Y. Li, "ReAct: Out-of-distribution detection with rectified activations," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 34, 2021, pp. 144–157.
[24] Y. Sun, Y. Ming, X. Zhu, and Y. Li, "Out-of-distribution detection with deep nearest neighbors," in *Proc. 39th Int. Conf. Mach. Learn. (ICML)*, PMLR, vol. 162, 2022, pp. 20827–20840.
[25] D. Hendrycks, M. Mazeika, and T. Dietterich, "Deep anomaly detection with outlier exposure," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019.
[26] M. Hein, M. Andriushchenko, and J. Bitterwolf, "Why ReLU networks yield high-confidence predictions far away from the training data and how to mitigate the problem," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 41–50.
[27] J. Bitterwolf, A. Meinke, and M. Hein, "Certifiably adversarially robust detection of out-of-distribution data," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, 2020, pp. 16085–16095.
[28] J. Yang et al., "OpenOOD: Benchmarking generalized out-of-distribution detection," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 35, 2022, pp. 32598–32611.
[29] J. Yang, K. Zhou, Y. Li, and Z. Liu, "Generalized out-of-distribution detection: A survey," *Int. J. Comput. Vis.*, vol. 132, no. 12, pp. 5635–5662, 2024.
[30] S. Fort, J. Ren, and B. Lakshminarayanan, "Exploring the limits of out-of-distribution detection," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 34, 2021, pp. 7068–7081.
[31] N. Carlini and D. Wagner, "Adversarial examples are not easily detected: Bypassing ten detection methods," in *Proc. 10th ACM Workshop Artif. Intell. Secur. (AISec)*, 2017, pp. 3–14.
[32] A. Meinke and M. Hein, "Towards neural networks that provably know when they don't know," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020.
[33] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry, "Robustness may be at odds with accuracy," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019.
[34] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, PMLR, vol. 97, 2019, pp. 7472–7482.
[35] J. Cohen, E. Rosenfeld, and Z. Kolter, "Certified adversarial robustness via randomized smoothing," in *Proc. 36th Int. Conf. Mach. Learn. (ICML)*, PMLR, vol. 97, 2019, pp. 1310–1320.
[36] W. M. John, "Constitutional AI: Scalable alignment through self-critique and revision," *Eduschool J. Artif. Intell. Res.*, vol. 1, no. 1, pp. 10–19, 2026, doi: 10.63090/EJAIR/3139.7336.0002.