Beyond AUROC & co. for evaluating out-of-distribution detection performance

Galadrielle Humblot-Renaux[1]   Sergio Escalera[1,2,3]   Thomas B. Moeslund[1]

[1]Visual Analysis and Perception lab, Aalborg University, Denmark
[2]Computer Vision Center, Universitat Autònoma de Barcelona, Spain
[3]Dept. of Mathematics and Informatics, Universitat de Barcelona, Spain

gegeh@create.aau.dk   sescalera@ub.edu   tbm@create.aau.dk

Abstract

While there has been a growing research interest in developing out-of-distribution (OOD) detection methods, there has been comparably little discussion around how these methods should be evaluated. Given their relevance for safe(r) AI, it is important to examine whether the basis for comparing OOD detection methods is consistent with practical needs. In this work, we take a closer look at the go-to metrics for evaluating OOD detection, and question the approach of exclusively reducing OOD detection to a binary classification task with little consideration for the detection threshold. We illustrate the limitations of current metrics (AUROC & its friends) and propose a new metric - Area Under the Threshold Curve (AUTC), which explicitly penalizes poor separation between ID and OOD samples. Scripts and data are available at https://github.com/glhr/beyond-auroc

1. Introduction

When deployed out in the wild, computer vision systems may be faced with image content which they simply are not equipped to handle. For instance, a model trained to recognize certain types of skin lesions, once deployed in clinical practice, may encounter images with a different kind of skin condition, or images with no lesions at all [5,20]. Thus, it is not enough for models to make accurate predictions on the kind of content that they were trained on - they should also be able to express whether a new input is familiar enough to make a reliable prediction.

The task of flagging images outside of a model's training domain is known as out-of-distribution (OOD) detection and is a growing line of research in computer vision [25, 30] with important implications for safe AI [1]. A broad range of methods have been proposed for equipping neural networks with OOD detection capabilities,
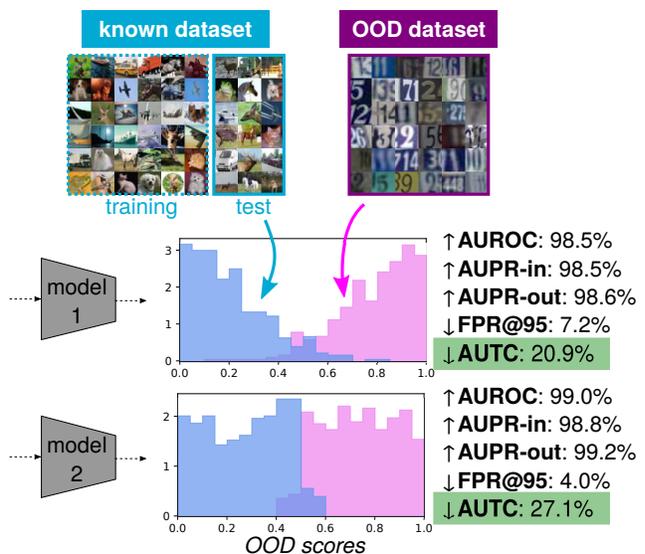

Figure 1. OOD evaluation setup illustrated with the CIFAR10 [10] vs. SVHN [23] pair. We visualize the OOD scores produced by 2 imaginary models as normalized histograms. Which model is better as an OOD detector? Popular metrics say model 2. We argue that more fine-grained metrics which take score distribution into account are needed for practical use, as they affect the choice of threshold for downstream decisions. We propose the AUTC metric which encourages separability between ID and OOD samples.

ranging from uncertainty quantification [11, 17], generative modelling [24], outlier exposure [4], to gradient-based [8], softmax-based [15], distance-based [26], or energy-based [19] approaches (among others). In this work, we abstract away from any specific OOD detection method, and rather focus on how OOD detection performance is quantitatively evaluated.

Fig. 1 illustrates our problem setup. We consider generic models which were trained on an image recognition dataset and, given an input image, output an OOD score alongside

their prediction - with a higher OOD score indicating that the image is more likely to be OOD. At evaluation time, the models are presented unseen in-distribution (ID) images (images within the training data distribution), along with OOD images from an unknown dataset. The scores are then aggregated and evaluated in terms of how well they can be used as a basis to distinguish between ID and OOD samples. This is the typical procedure in OOD benchmarks [30, 35]. We then question the exclusive use of binary classification metrics (AUROC, AUPR, FPR & co.) for quantitative comparison of different models, as they binarize OOD scores without considering their distribution or separability. Consider the example of Fig. 1, where *model 2* outperforms *model 1* across all standard metrics. Yet, looking at the distribution of scores, *model 1* may be preferable in practice, as it achieves a much clearer separation between ID vs. OOD samples, and allows more flexibility in the choice of threshold without drastic changes in detection performance. To expand on this intuition, we

- briefly review the status quo in terms of metrics for evaluating OOD detection and identify confusing discrepancies in their definition across the literature (Sec. 2)
- elaborate on some limitations of these metrics with the help of several illustrative examples (Sec. 3), and emphasize the need for a global detection threshold
- present an alternative view and new performance metric for evaluating OOD detectors with a focus on their downstream use, where separability between ID and OOD samples in terms of their OOD score is particularly important for the choice of a threshold (Sec. 4)

Related work Within computer vision, recent surveys have outlined the most commonly used performance metrics in OOD detection and related tasks [25, 31] - this aligns with our brief review. However, to the best of our knowledge, there is little to no existing work discussing their limitations or considering possible extensions.

Complementary to this paper, [29] discusses the design of OOD benchmarks for computer vision in terms of dataset splits, with the goal of minimizing semantic overlap between ID and OOD sets. [34] lays out practical guidelines and challenges for evaluating OOD detection when using medical data. In [5], the downstream implications of different kinds of mistakes are considered (e.g. flagging an OOD sample as ID vs. ID sample as OOD) and modelled as a cost matrix in terms of model trustworthiness. Within the context of visual question answering, [27] points out questionable but common practices in OOD benchmarks (e.g. tuning hyperparameters based on OOD performance), leading to misleadingly inflated performance.

Beyond the realm of OOD detection, [16] reviews some of the pitfalls of benchmark-oriented machine learning, a major one being the use of simplified metrics which do not capture important differences between methods. In a similar spirit to our work, [22] questions the accuracy metrics reported in metric learning, as they fail to capture a notion of class separation.

2. Performance metrics in OOD detection

Overall, the consensus is to treat OOD detection as a binary classification task, where the predicted continuous OOD score is binarized and compared to a true label (positive if the test sample is OOD, negative otherwise - or vice-versa). The prediction is then either considered a True Positive (TP), True Negative (TN), False Positive (FP) or False Negative (FN) - as visualized in Fig. 2. Following the seminal work in [7], the OOD detection literature has adopted the AUROC and AUPR as metrics of choice, thus bypassing the need to select a specific threshold. AUROC is often considered the main metric, and we did not find any works which did not report it. Alongside these, most works also report performance at a fixed detection threshold [4–6, 9, 13–15, 19, 21, 24, 30, 33, 36]. We briefly present each metric below.
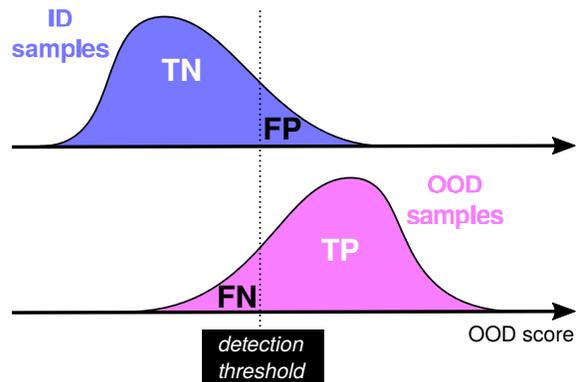

Figure 2. OOD scores are binarized based on a detection threshold.

Fixed-threshold metrics These consider performance at a specific operating point. The FPR@TPR metric measures the false positive rate for a given true positive rate - typically chosen to be 95% [9, 19, 30, 33], or sometimes 80% [24]. In a similar vein, some works instead report the TNR@95 (true negative rate at 95% TPR) [13, 14]. For an ideal detector, the TNR is 100% while the FPR is 0%.

Threshold-independent metrics These summarize OOD detection performance with a sliding threshold. The AUROC measures the area under the Receiver Operating Characteristic curve, obtained by plotting the TPR as a function

of the FPR. It can be interpreted as the probability that a positive sample is assigned a higher score than a negative sample - an AUROC of 100% indicates perfect separation, while an AUROC of 50% indicates full overlap (uninformative/random detector). The AUPR measures the area under the curve obtained by plotting recall (R) as a function of precision (P). Unlike AUROC, it is sensitive to sample size and the choice of positive class, hence it is common to distinguish between AUPR-in (ID samples considered positive) and AUPR-out (OOD samples considered positive) [15].

Detection error/accuracy Some works [13] report this as the best detection accuracy/error across all possible thresholds, while others report it for a fixed TP rate (e.g. 95% in [15]). As discussed in [7], a drawback of using detection accuracy as a metric is that it is skewed by class imbalance (ie. it assumes an equal number of ID and OOD samples).

Positive or negative, that is the question We note a mismatch in the literature in terms of which class (ID or OOD) is considered positive or negative for the computation of metrics. While one set of papers treat ID samples as the positive class [7, 13, 15, 30], another instead treats OOD samples as positive [9, 21, 28]. Some works fail to mention this definition altogether [17]. While this may only seem like a minor difference of terminology, it affects the AUPR and FPR computation - both widely used metrics. For instance, the FPR@95 metric reported in the recent benchmark [30]

vs. the one reported in [28] are in fact different metrics due to this mismatch in class definition. Special care should therefore be taken to avoid such inconsistencies. In the context of this paper, we consider OOD samples to be positive unless otherwise specified.

3. What's the problem?

3.1. Let histograms (s)peak for themselves

Assessing the performance of OOD detection methods purely in terms of binary classification metrics reduces the evaluation to a comparison of ratios between the number of TP, TN, FP and FN predictions, without considering how OOD scores are distributed. That is, two models with the same performance may differ widely in terms of how clearly they separate ID from OOD samples. In a similar spirit to Fig. 1, we illustrate this effect in Fig. 3, where we simulate OOD scores produced by 9 imaginary models (which were trained and evaluated on the same imaginary data), such that models on a given row exhibit near-identical performance in terms of AUROC. Other standard metrics for each model are reported in Fig. 4.

Is high performance all we need? The first row in Fig. 3 shows examples of "perfect" OOD detection models in terms of standard metrics, with no overlap between the two classes (such high detection performance is not unheard of - for instance, the method in [2] achieves 100% AUROC and
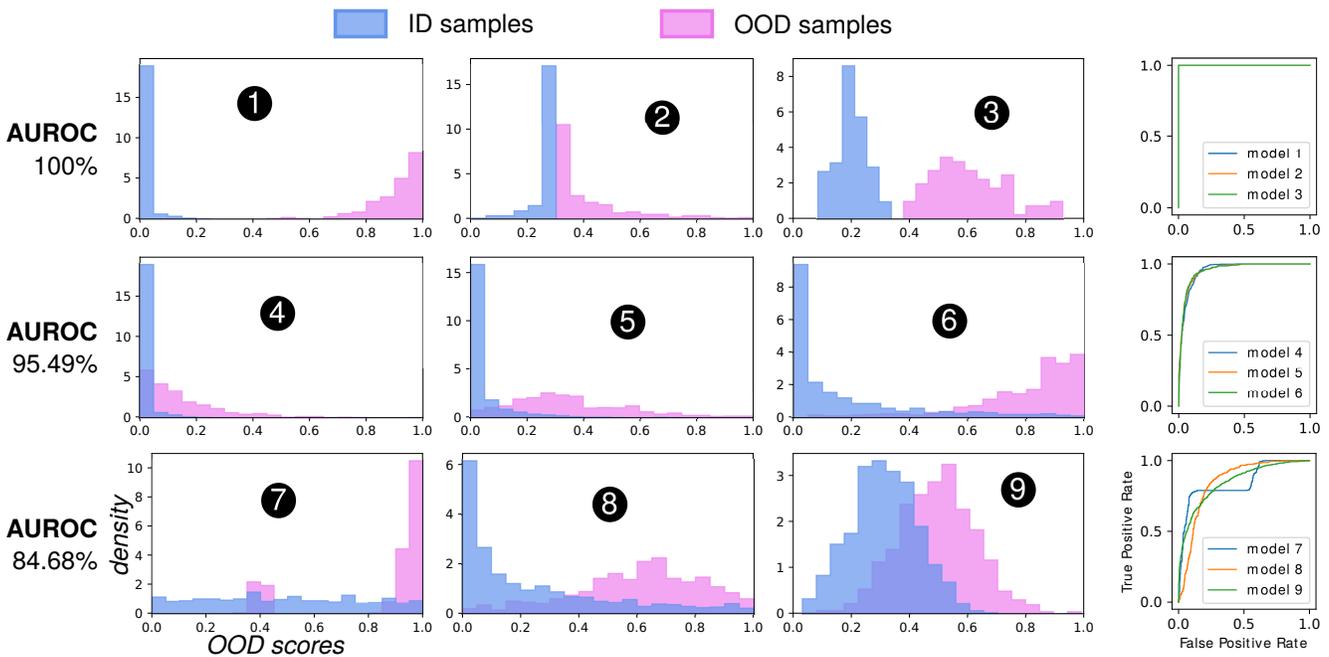

Figure 3. Histograms (normalized) of OOD scores for ID vs. OOD samples produced by 9 imaginary models. The models on a given row achieve the same (rounded to 4 decimal places) OOD performance in terms of AUROC. On the right are ROC curves.
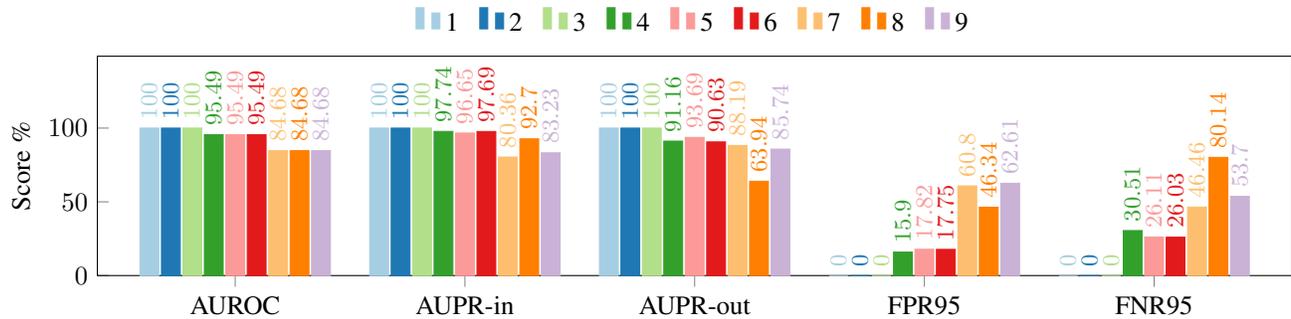
Figure 4. Performance of the 9 imaginary models in terms of standard metrics. Models are numbered according to Fig. 3.

AUPR on the CIFAR10/100 vs. SVHN pair). Despite their identical performance, if tasked with picking a model to deploy in a practical application, one would most likely prefer model 1 over model 2 due to the clear separation between ID and OOD samples. Model 2 inspires less confidence outside of a benchmark scenario (where we only consider images present in the test set), as its performance is extremely sensitive to the choice of threshold. Yet, none of the standard metrics capture this distinction.

In the second row, model 4 is a prime example of a model achieving what is considered an "excellent" OOD detection performance by common standards [7], but very poor separation between scores assigned to ID vs. OOD samples. Trying to select a suitable operating point for this model would be less straightforward than for model 6 - if set slightly too low, many inputs would be wrongly flagged as OOD, but if slightly higher, a large portion of OOD samples would be missed. Across the performance metrics in Fig. 4, FNR@95 is the only metric penalizing model 4.

Lastly, the models in the third row would be considered "good" OOD detectors based on their AUROC of almost 85% [7]. Model 7 exhibits quite undesirable behaviour, with OOD scores for ID samples uniformly distributed across the whole range, yet it achieves the best performance in terms of AUPR-out and FNR@95. Models 8 and 9 have a clearer separation between ID vs. OOD samples, with a sensible threshold lying around 0.4. The large differences between AUPR-in vs. AUPR-out and FPR@95 vs. FNR@95 results for model 8 in Fig. 4 also highlights that reporting only one "side" of these metrics can give a misleading picture of performance.

In a nutshell As we have shown with the examples of Fig. 3, standard metrics are sensitive to the amount of *overlap* between scores assigned to ID vs. OOD samples, but are blind to the level of *separation* between them. Indeed, a model achieving perfect performance in terms of AUROC or AUPR only means that there exists at least *one* threshold for which the FPR and FNR are 0. We argue that having a wide range of sensible thresholds to choose from is a desir-
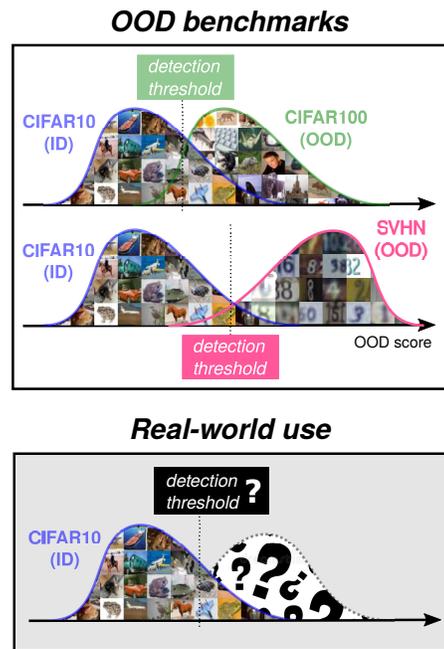

Figure 5. In a benchmark setting, once the model has been trained (on CIFAR10 in this example), the threshold for metrics is often set based on performance on each individual OOD dataset (e.g. CIFAR100 or SVHN). However, deployment of the model in a practical setting requires the choice of a single sensible threshold which is not tailored to a specific OOD dataset.

able property for an OOD detector.

3.2. About that threshold

Besides the fact that the metrics themselves only tell part of the story, another practical concern is that current evaluation of OOD detectors essentially considers each ID vs. OOD dataset pair as its own classification task. For example, a model trained on CIFAR10, will then independently be evaluated on CIFAR10 vs. SVHN and CIFAR10 vs. CIFAR100 (and often several other OOD datasets [21]), with the above-mentioned metrics reported for each pair.

Thus, even when employing a fixed-threshold metric, the threshold may vary across the different OOD datasets in the evaluation - as illustrated in Fig. 5. We argue that this approach is fundamentally misaligned with a real-use setting, for which a single detection threshold has to be chosen.

4. What next?

We suggest a shift of perspective: consider the OOD scores not just as the output of a classifier, but as an *input* to a decision function which has to determine whether to discard the model's prediction on the original task. The selection of a threshold for this decision function will determine the actual OOD detection performance for new images, and is therefore a safety-critical design choice.

With the threshold in mind, we present a new performance metric along with some recommendations for the evaluation of OOD detectors.

4.1. Enter a new metric

If model performance at evaluation time (that is, for a single ID vs. OOD dataset pair) is extremely sensitive to the choice of threshold, then it is difficult to imagine robust performance at run-time, where the diversity of OOD samples is expected to increase. Yet, standard area-based performance metrics do not capture the relation between a change of threshold and a change in performance.

To address this gap and quantify important differences between models which are not caught by existing metrics, we therefore propose a new metric, the Area Under the Threshold Curve (AUTC), which explicitly penalizes poor separation between ID and OOD samples. Increased sep-

arability is sometimes mentioned as a desired property for OOD detectors in the state of the art [15], but has not been previously quantified.

In contrast with the ROC curve or PR curve, our metric is based on a visualization which explicitly shows the effect of the threshold on OOD detection performance by plotting the FPR and FNR of the detector as a function of the detection threshold. As shown in Fig. 6, not only are these plots convenient for visually choosing a threshold, they also reveal stark differences between models - even those with perfect AUROC - in terms of how quickly the performance degrades when moving away from the curves' crossing point. They also implicitly combine both a measure of performance and separability: the lower the crossing point, the higher the OOD detection performance with a good choice of threshold, and the closer the curves are glued to the Y axis on each side, the more ID and OOD scores are concentrated at opposite sides of the score ranges.

Notably, as the area under the FPR curve (AUFPR) and under the FNR curve (AUFNR) shrink, we approach an ideal detector which assigns a score of 0 to all ID samples, and 1 to all OOD samples.

AUTC metric   We summarize these curves as a single metric by combining the AUFPR and AUFNR, and averaging them to obtain a single value within $[0, 1]$:

$$\text{AUTC} = \frac{\text{AUFPR} + \text{AUFNR}}{2}$$

In practice, the areas can be computed via the trapezoidal rule, as is commonly done for the AUROC or AUPR.
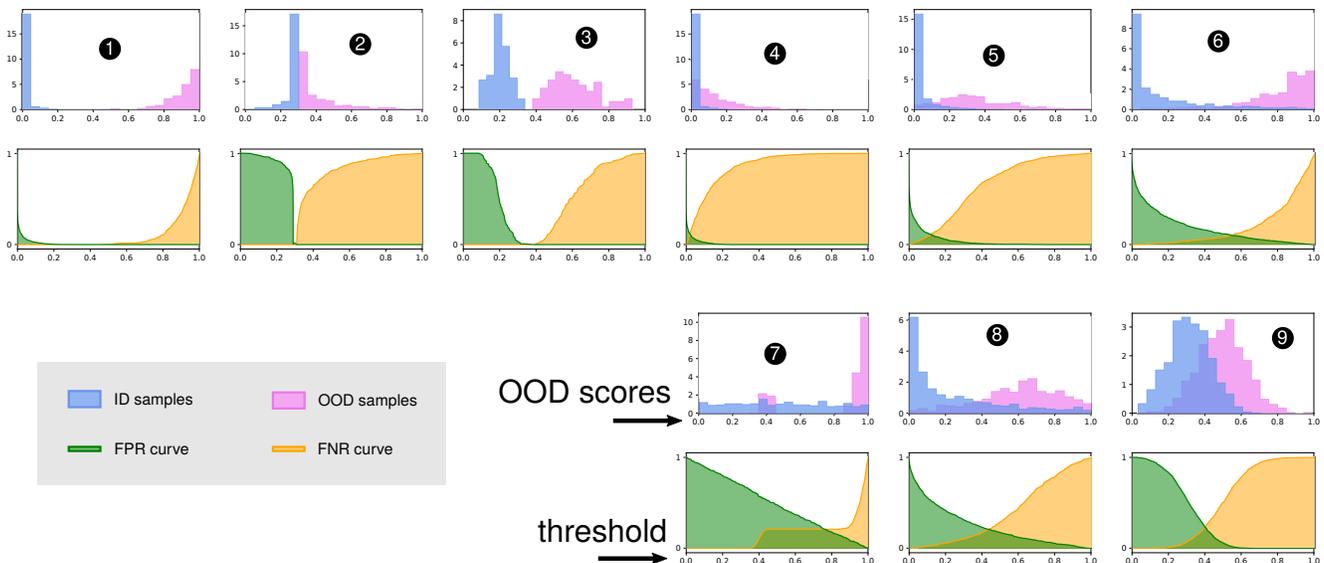

Figure 6. Visualization of the FPR and FNR vs. threshold curves used to compute our proposed metric. The corresponding histogram of each model is shown above them for reference.
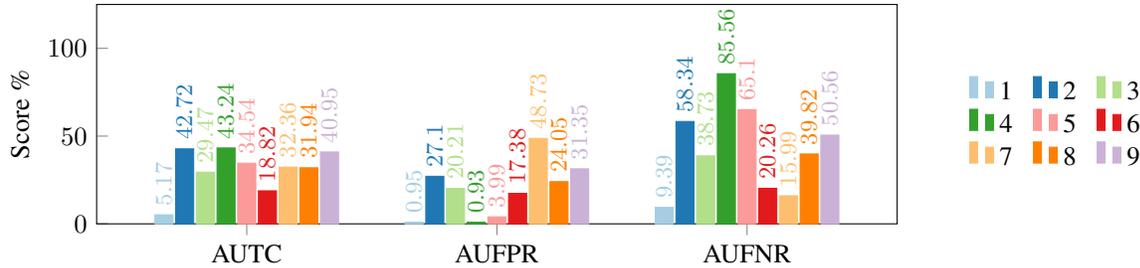
Figure 7. Our proposed OOD performance metrics for the same 9 models. Lower performance is better.

Some properties of the AUTC:

- a lower value is better: it is equal to 1 for the worst detector with complete separation, 0.5 for a random detector (no separation), and 0 for a perfect detector with complete separation.
- it is not sensitive to sample size or to the definition of positive vs. negative class.
- it assumes OOD scores between 0 and 1.

In Fig. 7, we show the AUTC computed for the 9 models, as well as the corresponding AUFPR and AUFNR. This comparison reveals significant differences compared to the standard metrics from Fig. 4. This time, in terms of AUTC, Models 1 and 6 clearly stand out compared to the other models, as the metric encourages strong separability, while model 4 obtains the worst score despite its AUROC of over 95%. Looking at the AUFPR and AUFNR allows us to separately quantify the spread and distance from 0/1 for ID vs. OOD samples. For instance, Models 4-6 have a significantly lower AUFPR than Models 7-8, as the scores assigned to ID samples are more concentrated around 0.

4.2. Other considerations for evaluation

We note several limitations of the AUTC metric:

1. The AUTC encourages strong separability - that is, it encourages OOD scores to be concentrated around 0 for all ID samples, and 1 for all OOD samples. This behaviour may not be desirable if one instead wishes the OOD scores to be well-calibrated (that is, for the OOD score to be a good indicator of the *probability* of a sample being OOD).

2. Since it simply averages the AUFPR and AUFNR, the AUTC gives equal weight to false negatives and false positives. In practice, the cost of these two types of mistakes may not be symmetrical (as discussed in [5]). We therefore recommend separately reporting the AUFPR and AUFNR, and/or giving them different weights in the AUTC computation based on severity.

3. As is the case for other curves parametrized by a threshold [3], the AUTC is sensitive to transformations of the OOD scores.

Furthermore, much like AUROC and AUPR, our proposed metric is a summary metric covering all possible thresholds. When comparing OOD detection methods, it should be accompanied by a fixed-threshold metric (e.g. by reporting FPR and FNR at a specific operating point). However, contrary to common practice, we emphasize that this fixed threshold should not be tailored to the OOD datasets used in the final evaluation. It should instead be tuned either on the ID dataset, or on a separate validation OOD set which is not used during final evaluation - as we cannot assume to know the distribution of unseen OOD samples. During evaluation, to reflect real-world conditions, the same threshold should be used across all OOD datasets when reporting fixed-threshold metrics.

5. A concrete example

Moving away from imaginary models and synthetically-generated data, we demonstrate our approach on real OOD models and ID vs. OOD dataset pairs. We select 2 models from the state of the art trained on the good old CIFAR10 [10] dataset, and compare their OOD detection performance on multiple unseen datasets (CIFAR100, tinyImageNet [12], SVHN [23], and LSUN [32]). We briefly present the models below, and refer to the original papers for details:

1. Out-of-DIstribution detector for Neural networks (ODIN) from [15] applies temperature scaling and input perturbations to a pre-trained neural network. The OOD score is based on the maximum softmax probability. We use the DenseNet model weights from the official code repository[1].

2. Spectral-normalized Neural Gaussian Process (SNGP) from [18] combines a distance-preserving feature extractor with an approximate Gaussian process as output layer. The OOD score is taken as the Dempster-Shafer metric. We train a model following a third-party PyTorch implementation[2].

---

[1]https://github.com/facebookresearch/odin
[2]https://github.com/y0ast/DUE

(a) ODIN

		OOD datasets		
		tiny Imagenet	LSUN	SVHN
	AUROC ↑	99.11	97.89	89.90
	AUFPR ↓	——— 45.17 ———		
	AUFNR ↓	21.97	25.24	35.11
	AUTC ↓	**33.57**	**35.21**	**40.41**
@test	FPR ↓ / FNR ↓	4.71	4.40	18.51
@95TNR	FPR ↓	——— 5.00 ———		
	FNR ↓	4.30	11.40	50.92
@val	FPR ↓	——— 19.56 ———		
	FNR ↓	0.46	1.88	17.00

(b) SNGP

		OOD datasets		
		tiny Imagenet	LSUN	SVHN
	AUROC ↑	97.38	99.09	99.06
	AUFPR ↓	——— 2.85 ———		
	AUFNR ↓	49.98	29.41	33.79
	AUTC ↓	**26.42**	**16.13**	**18.32**
@test	FPR ↓ / FNR ↓	8.36	5.01	4.97
@95TNR	FPR ↓	——— 5.00 ———		
	FNR ↓	14.11	5.02	4.95
@val	FPR ↓	——— 16.22 ———		
	FNR ↓	3.19	0.89	0.77

Table 1. Quantitative comparison of the 2 models' OOD detection performance. Scores are in percentages, and the proposed AUTC metric is highlighted in bold. Note that when fixing a global threshold, the FPR is the same across all OOD datasets as it only depends on the distribution of OOD scores for the ID data (CIFAR10).

Note that the purpose of this experiment is not to pit one method against another, but rather to show our proposed metric and evaluation procedure in action.

More plots We use CIFAR100 as a validation dataset for the detection threshold (as it is "closer" to the training set in terms of appearance than the others), and the rest of the OOD datasets for the final evaluation. As visualized in Fig. 8, although they exhibit comparable detection performance (similar amount of overlap between the FPR and FNR curves), the models differ widely in terms of how their OOD scores are distributed for ID vs. OOD samples. For ODIN, a reasonable threshold lies around 0.55, while SNGP concentrates its scores for ID samples around 0. As the threshold increases from the crossing point, the FNR increases more rapidly for ODIN than SNGP.

Finally a table We summarize the quantitative results in Tab. 1. For each model and ID vs. OOD pair, we report the AUROC (main standard metric), AUTC (our proposed metric) as threshold-independent performance metrics. We also measure the FNR (probably of misclassifying an OOD sample as ID) for several fixed thresholds:

- **@test** - the point at which the FPR and FNR on the OOD set are equal (this threshold is specific to each OOD dataset). We include this as reference, to show ideal performance.
- **@95TNR** - the point at which the TNR is at least 95%. This threshold only depends on the distribution of ID scores (thus stays the same across OOD datasets).
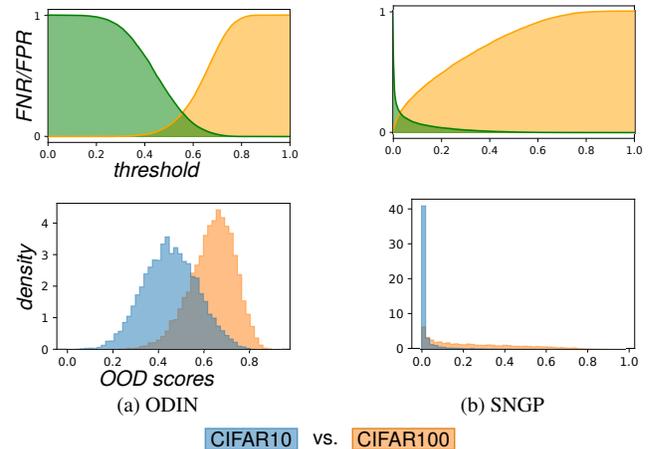

Figure 8. FPR (green) & FNR (yellow) curves and normalized histograms of OOD scores predicted by the 2 models on CIFAR10 (ID dataset) vs. CIFAR100 (val OOD dataset).

- **@val** - the point at which the FPR and FNR on the validation dataset CIFAR100 are equal (also stays the same across OOD datasets).

Looking at the threshold-independent metrics (AUROC to AUTC in Tab. 1), both models achieve similar and high (potential for) OOD detection performance, except for SVHN where there is a difference of 10 percentage points in AUROC between ODIN and SNGP. Our AUTC metric correlates with AUROC performance for a given model, while also indicating that SNGP produces better separability between ID vs. OOD samples. Note that the AUFPR is constant across OOD datasets as it only depends on the distri-

bution of ID samples - SNGP has a much lower AUFPR due to a strong concentration of OOD scores around 0 for ID samples.

The threshold-specific results (FPR and FNR in Tab. 1) show how significantly the choice of threshold can impact performance. The performance at @test is unrealistic in practice, as it assumes that the threshold can be adjusted for each OOD dataset - as shown in Fig. 9. Fixing a global threshold based on the ID dataset scores (@95TNR) or the validation OOD set (@val) widens the performance gap across different ID vs. OOD pairs. When shifting the global threshold from one to the other, we see larger fluctuations in performance on OOD datasets with a higher AUTC (SVHN for ODIN, and tinyImagenet for SNGP).

(a) ODIN   (b) SNGP

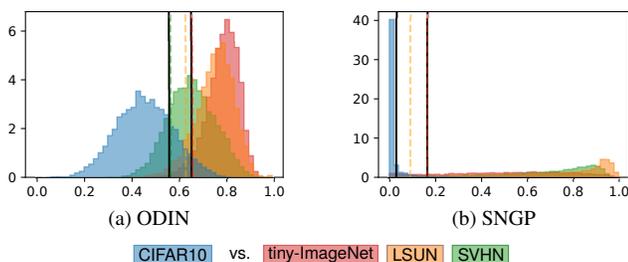CIFAR10   vs.   tiny-ImageNet   LSUN   SVHN

Figure 9. Normalized histograms of OOD scores for the test OOD pairs. The global (black) and @test (dashed, colored by dataset) thresholds are shown as vertical lines.

6. Zooming out

Beyond CIFAR Throughout the illustrative examples of this paper, we have considered the context of image classification, as this is the most common setting in OOD benchmarks for computer vision [30]. However, our analysis of OOD metrics extends well beyond this setting, as it abstracts away from any particular input modality (images, audio, point clouds...) or main task (classification, regression..), as well as how or at what level of granularity (e.g. image-level, pixel-level...) the OOD scores are produced.

Future work We have emphasized the need for evaluating OOD detection with the choice of a global threshold in mind, as this choice would have to made for any practical application: samples with an OOD score above this threshold are flagged as OOD, allowing the system to fallback to a safe strategy (e.g. requesting human input) rather than allowing the model to make a prediction. Investigating how a "good" global threshold should be chosen (without assuming the distributions of OOD datasets are known) and whether ID vs. OOD separability indeed translates to more robust OOD detection performance are important directions for future research, as well as developing methods which incorporate the choice of threshold in the model design itself.

7. Conclusion

In this work, we have focused on the quantitative evaluation of OOD detectors, highlighting that current performance metrics can lead to misleading comparisons between methods due to a terminology mismatch in the OOD detection literature, and can obfuscate some important differences between models such as their ability to produce clearly separated scores for ID vs. OOD samples. With concrete examples, we have shown that achieving a high performance in terms of AUROC is only the first step towards utilizing OOD detection in practical settings, and that the choice of a detection threshold should be treated as an important hyperparameter rather than an afterthought. We have presented a new metric which can serve as a complementary basis for comparing OOD detection models in terms of how well they separate ID from OOD samples by OOD score. We hope that this paper serves as a starting point to encourage further discussion around how OOD detection methods should be evaluated to align with the goals of practical and safe AI.

8. Acknowledgements

References

[1] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016. 1

[2] Senqi Cao and Zhongfei Zhang. Deep hybrid models for out-of-distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4733–4743, June 2022. 3

[3] Akshay Raj Dhamija, Manuel Günther, and Terrance E. Boult. Reducing network agnostophobia. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 9175–9186, Red Hook, NY, USA, 2018. Curran Associates Inc. 6

[4] Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 7068–7081. Curran Associates, Inc., 2021. 1, 2

[5] Abhijit Guha Roy, Jie Ren, Shekoofeh Azizi, Aaron Loh, Vivek Natarajan, Basil Mustafa, Nick Pawlowski, Jan Freyberg, Yuan Liu, Zach Beaver, Nam Vo, Peggy Bui, Samantha Winter, Patricia MacWilliams, Greg S. Corrado, Umesh Telang, Yun Liu, Taylan Cemgil, Alan Karthikesalingam,

Balaji Lakshminarayanan, and Jim Winkens. Does your dermatology classifier know what it doesn't know? detecting the long-tail of unseen conditions. *Medical Image Analysis*, 75:102274, 2022. 1, 2, 6

[6] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joseph Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 8759–8773. PMLR, 17–23 Jul 2022. 2

[7] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017. 2, 3, 4

[8] Rui Huang, Andrew Geng, and Yixuan Li. On the importance of gradients for detecting distributional shifts in the wild. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 677–689. Curran Associates, Inc., 2021. 1

[9] R. Huang and Y. Li. Mos: Towards scaling out-of-distribution detection for large semantic space. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8706–8715, Los Alamitos, CA, USA, jun 2021. 2, 3

[10] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, University of Toronto, Ontario, 2009. 1, 6

[11] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 1

[12] Ya Le and Xuan S. Yang. Tiny imagenet visual recognition challenge. Technical report, Stanford University, 2015. 6

[13] Kimin Lee, Honglak Lee, Kibok Lee, and Jinwoo Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. In *International Conference on Learning Representations*, 2018. 2, 3

[14] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. 2

[15] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. In *International Conference on Learning Representations*, 2018. 1, 2, 3, 5, 6

[16] Thomas Liao, Rohan Taori, Deborah Raji, and Ludwig Schmidt. Are we learning yet? a meta review of evaluation failures across machine learning. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran Associates, Inc., 2021. 2

[17] Jeremiah Liu, Zi Lin, Shreyas Padhy, Dustin Tran, Tania Bedrax Weiss, and Balaji Lakshminarayanan. Simple and principled uncertainty estimation with deterministic deep learning via distance awareness. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 7498–7512. Curran Associates, Inc., 2020. 1, 3

[18] Jeremiah Zhe Liu, Shreyas Padhy, Jie Ren, Zi Lin, Yeming Wen, Ghassen Jerfel, Zachary Nado, Jasper Snoek, Dustin Tran, and Balaji Lakshminarayanan. A simple approach to improve single-model deep uncertainty via distance-awareness. *Journal of Machine Learning Research*, 24(42):1–63, 2023. 6

[19] Weitang Liu, Xiaoyun Wang, John Owens, and Yixuan Li. Energy-based out-of-distribution detection. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 21464–21475. Curran Associates, Inc., 2020. 1, 2

[20] Deval Mehta, Yaniv Gal, Adrian Bowling, Paul Bonnington, and Zongyuan Ge. Out-of-distribution detection for long-tailed and fine-grained skin lesion images. In Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, pages 732–742, Cham, 2022. Springer Nature Switzerland. 1

[21] Sina Mohseni, Mandar Pitale, JBS Yadawa, and Zhangyang Wang. Self-supervised learning for generalizable out-of-distribution detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):5216–5223, Apr. 2020. 2, 3, 4

[22] Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 681–699, Cham, 2020. Springer International Publishing. 2

[23] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading Digits in Natural Images with Unsupervised Feature Learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011. 1, 6

[24] 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. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 1, 2

[25] 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. *Transactions on Machine Learning Research*, 2022. 1, 2

[26] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In Ka-

malika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 20827–20840. PMLR, 17–23 Jul 2022. 1

[27] Damien Teney, Ehsan Abbasnejad, Kushal Kafle, Robik Shrestha, Christopher Kanan, and Anton van den Hengel. On the value of out-of-distribution testing: An example of goodhart's law. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 407–417. Curran Associates, Inc., 2020. 2

[28] Haoqi Wang, Zhizhong Li, Litong Feng, and Wayne Zhang. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4921–4930, June 2022. 3

[29] Jingkang Yang, Haoqi Wang, Litong Feng, Xiaopeng Yan, Huabin Zheng, Wayne Zhang, and Ziwei Liu. Semantically coherent out-of-distribution detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8301–8309, October 2021. 2

[30] 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. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 32598–32611. Curran Associates, Inc., 2022. 1, 2, 3, 8

[31] Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *arXiv preprint arXiv:2110.11334*, 2021. 2

[32] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2016. 6

[33] Qing Yu and Kiyoharu Aizawa. Unsupervised out-of-distribution detection by maximum classifier discrepancy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2

[34] Karina Zadorozhny, Patrick Thoral, Paul Elbers, and Giovanni Cinà. *Multimodal AI in Healthcare: A Paradigm Shift in Health Intelligence*, chapter Out-of-Distribution Detection for Medical Applications: Guidelines for Practical Evaluation, pages 137–153. Springer International Publishing, Cham, 2023. 2

[35] David Zimmerer, Peter M. Full, Fabian Isensee, Paul Jäger, Tim Adler, Jens Petersen, Gregor Köhler, Tobias Ross, Annika Reinke, Antanas Kascenas, Bjørn Sand Jensen, Alison Q. O'Neil, Jeremy Tan, Benjamin Hou, James Batten, Huaqi Qiu, Bernhard Kainz, Nina Shvetsova, Irina Fedulova, Dmitry V. Dylov, Baolun Yu, Jianyang Zhai, Jingtao Hu, Runxuan Si, Sihang Zhou, Siqi Wang, Xinyang Li, Xuerun Chen, Yang Zhao, Sergio Naval Marimont, Giacomo Tarroni, Victor Saase, Lena Maier-Hein, and Klaus Maier-Hein. Mood 2020: A public benchmark for out-of-distribution detection and localization on medical images. *IEEE Transactions on Medical Imaging*, 41(10):2728–2738, 2022. 2

[36] Ev Zisselman and Aviv Tamar. Deep residual flow for out of distribution detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2