Augmenting Softmax Information for Selective Classification with Out-of-Distribution Data

Guoxuan Xia and Christos-Savvas Bouganis

Imperial College London
{g.xia21,christos-savvas.bouganis}@imperial.ac.uk

Abstract. Detecting out-of-distribution (OOD) data is a task that is receiving an increasing amount of research attention in the domain of deep learning for computer vision. However, the performance of detection methods is generally evaluated on the task in isolation, rather than also considering potential downstream tasks in tandem. In this work, we examine selective classification in the presence of OOD data (SCOD). That is to say, the motivation for detecting OOD samples is to reject them so their impact on the quality of predictions is reduced. We show under this task specification, that existing post-hoc methods perform quite differently compared to when evaluated only on OOD detection. This is because it is no longer an issue to conflate in-distribution (ID) data with OOD data *if the ID data is going to be misclassified*. However, the conflation within ID data of correct and incorrect predictions becomes undesirable. We also propose a novel method for SCOD, Softmax Information Retaining Combination (SIRC), that augments softmax-based confidence scores with feature-agnostic information such that their ability to identify OOD samples is improved without sacrificing separation between correct and incorrect ID predictions. Experiments on a wide variety of ImageNet-scale datasets and convolutional neural network architectures show that SIRC is able to consistently match or outperform the baseline for SCOD, whilst existing OOD detection methods fail to do so. Code is available at `https://github.com/Guoxoug/SIRC`.

1 Introduction

Out-of-distribution (OOD) detection [49], i.e. identifying data samples that do not belong to the training distribution, is a task that is receiving an increasing amount of attention in the domain of deep learning [4, 6, 15, 16, 19, 22, 31–33, 39, 41, 45, 46, 48–50]. The task is often motivated by safety-critical applications, such as healthcare and autonomous driving, where there may be a large cost associated with sending a prediction on OOD data downstream.

However, in spite of a plethora of existing research, there is generally a lack of focus with regards to the specific motivation behind OOD detection in the literature, other than it is often done as part of the pipeline of another primary task, e.g. image classification. As such the task is evaluated in isolation and formulated as binary classification between in-distribution (ID) and OOD data. In

this work we consider the question *why exactly do we want to do OOD detection during deployment?* We focus on the problem setting where the primary objective is classification, and we are motivated to detect and then reject OOD data, as predictions on those samples will incur a cost. That is to say the task is selective classification [5, 8] where OOD data has polluted the input samples. Kim et al. [27] term this problem setting *unknown detection*. However, we prefer to use Selective Classification in the presence of Out-of-Distribution data (SCOD) as we would like to emphasise the downstream classifier as the objective, and will refer to the task as such in the remainder of the paper.

The *key difference* between this problem setting and OOD detection is that *both* OOD data *and* incorrect predictions on ID data will incur a cost [27]. It does not matter if we reject an ID sample if it would be incorrectly classified anyway. As such we can view the task as separating correctly predicted ID samples (ID✓) from misclassified ID samples (ID✗) and OOD samples. This reveals a potential blind spot in designing approaches solely for OOD detection, as the cost of ID misclassifications is ignored. The *key contributions* of this work are:

1. Building on initial results from [27] that show poor SCOD performance for existing methods designed for OOD detection, we show novel insight into the behaviour of different post-hoc (after-training) detection methods for the task of SCOD. Improved OOD detection often comes directly at the expense of SCOD performance. Moreover, the relative SCOD performance of different methods varies with the proportion of OOD data found in the test distribution, the relative cost of accepting ID✗ vs OOD, as well as the distribution from which the OOD data samples are drawn.
2. We propose a novel method, targeting SCOD, Softmax Information Retaining Combination (SIRC), that aims to improve the OOD|ID✓ separation of softmax-based methods, whilst retaining their ability to identify ID✗. It consistently outperforms or matches the baseline maximum softmax probability (MSP) approach over a wide variety of OOD datasets and convolutional neural network (CNN) architectures, unlike existing OOD detection methods.

2 Preliminaries

Neural Network Classifier For a K-class classification problem we learn the parameters $\boldsymbol{\theta}$ of a discriminative model $P(y|\boldsymbol{x};\boldsymbol{\theta})$ over labels $y \in \mathcal{Y} = \{\omega_k\}_{k=1}^{K}$ given inputs $\boldsymbol{x} \in \mathcal{X} = \mathbb{R}^D$, using finite training dataset $\mathcal{D}_{\text{tr}} = \{y^{(n)}, \boldsymbol{x}^{(n)}\}_{n=1}^{N}$ sampled independently from true joint data distribution $p_{\text{tr}}(y, \boldsymbol{x})$. This is done in order to make predictions $\hat{y}$ given new inputs $\boldsymbol{x}^* \sim p_{\text{tr}}(\boldsymbol{x})$ with unknown labels,

$$\hat{y} = f(\boldsymbol{x}^*) = \arg\max_{\omega} P(\omega|\boldsymbol{x}^*;\boldsymbol{\theta}) \ , \tag{1}$$

where f refers to the classifier function. In our case, the parameters $\boldsymbol{\theta}$ belong to a deep neural network with categorical softmax output $\boldsymbol{\pi} \in [0,1]^K$,

$$P(\omega_i|\boldsymbol{x};\boldsymbol{\theta}) = \pi_i(\boldsymbol{x};\boldsymbol{\theta}) = \exp v_i(\boldsymbol{x}) / \sum_{k=1}^{K} \exp v_k(\boldsymbol{x}) \ , \tag{2}$$

where the logits $\boldsymbol{v} = \boldsymbol{W}\boldsymbol{z}+\boldsymbol{b}$ $(\in \mathbb{R}^K)$ are the output of the final fully-connected layer with weights $\boldsymbol{W} \in \mathbb{R}^{K\times L}$, bias $\boldsymbol{b} \in \mathbb{R}^K$, and final hidden layer features $\boldsymbol{z} \in \mathbb{R}^L$ as inputs. Typically $\boldsymbol{\theta}$ are learnt by minimising the cross entropy loss, such that the model approximates the true conditional distribution $P_{\text{tr}}(y|\boldsymbol{x})$,

$$\begin{aligned}\mathcal{L}_{\text{CE}}(\boldsymbol{\theta}) &= -\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}\delta(y^{(n)},\omega_k)\log P(\omega_k|\boldsymbol{x}^{(n)};\boldsymbol{\theta}) \\ &\approx -\mathbb{E}_{p_{\text{tr}}(\boldsymbol{x})}\left[\sum_{k=1}^{K}P_{\text{tr}}(\omega_k|\boldsymbol{x})\log P(\omega_k|\boldsymbol{x};\boldsymbol{\theta})\right] = \mathbb{E}_{p_{\text{tr}}}\left[KL\left[P_{\text{tr}}||P_{\boldsymbol{\theta}}\right]\right] + A\ ,\end{aligned} \tag{3}$$

where $\delta(\cdot,\cdot)$ is the Kronecker delta, A is a constant with respect to $\boldsymbol{\theta}$ and $\text{KL}[\cdot,\cdot]$ is the Kullback–Leibler divergence.

Selective Classification A selective classifier [5] can be formulated as a pair of functions, the aforementioned classifier $f(\boldsymbol{x})$ (in our case given by Eq. 1) that produces a prediction $\hat{y}$, and a binary rejection function

$$g(\boldsymbol{x};t) = \begin{cases} 0 \text{ (reject prediction)}, & \text{if } S(\boldsymbol{x}) < t \\ 1 \text{ (accept prediction)}, & \text{if } S(\boldsymbol{x}) \geq t\ , \end{cases} \tag{4}$$

where t is an operating threshold and S is a scoring function which is typically a measure of predictive confidence (or $-S$ measures uncertainty). Intuitively, a selective classifier chooses to reject if it is uncertain about a prediction.

Problem Setting We consider a scenario where, during deployment, classifier inputs $\boldsymbol{x}^*$ may be drawn from either the training distribution $p_{\text{tr}}(\boldsymbol{x})$ (ID) or another distribution $p_{\text{OOD}}(\boldsymbol{x})$ (OOD). That is to say,

$$\boldsymbol{x}^* \sim p_{\text{mix}}(\boldsymbol{x}), \quad p_{\text{mix}}(\boldsymbol{x}) = \alpha p_{\text{tr}}(\boldsymbol{x}) + (1-\alpha)p_{\text{OOD}}(\boldsymbol{x})\ , \tag{5}$$

where $\alpha \in [0,1]$ reflects the proportion of ID to OOD data found in the wild. Here "Out-of-Distribution" inputs are defined as those drawn from a distribution with label space that does not intersect with the training label space $\mathcal{Y}$ [49]. For example, an image of a car is considered OOD for a CNN classifier trained to discriminate between different types of pets.

We now define the predictive loss on an accepted sample as

$$\mathcal{L}_{\text{pred}}(f(\boldsymbol{x}^*)) = \begin{cases} 0, & \text{if } f(\boldsymbol{x}^*) = y^*, \quad y^*, \boldsymbol{x}^* \sim p_{\text{tr}}(y,\boldsymbol{x}) \quad (\text{ID}\checkmark) \\ \beta, & \text{if } f(\boldsymbol{x}^*) \neq y^*, \quad y^*, \boldsymbol{x}^* \sim p_{\text{tr}}(y,\boldsymbol{x}) \quad (\text{ID}\times) \\ 1-\beta, & \text{if } \boldsymbol{x}^* \sim p_{\text{OOD}}(\boldsymbol{x}) \quad (\text{OOD})\ , \end{cases} \tag{6}$$

where $\beta \in [0,1]$, and define the selective risk as in [8],

$$R(f,g;t) = \frac{\mathbb{E}_{p_{\text{mix}}(\boldsymbol{x})}[g(\boldsymbol{x};t)\mathcal{L}_{\text{pred}}(f(\boldsymbol{x}))]}{\mathbb{E}_{p_{\text{mix}}(\boldsymbol{x})}[g(\boldsymbol{x};t)]}\ , \tag{7}$$

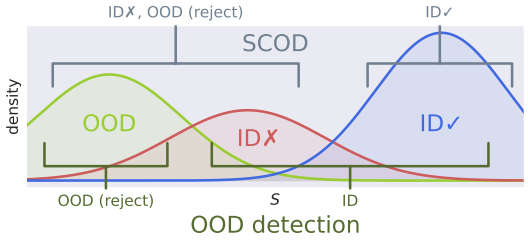

Fig. 1. Illustrative sketch showing how SCOD differs to OOD detection. Densities of OOD samples, misclassifications (ID✗) and correct predictions (ID✓) are shown with respect to confidence score S. For OOD detection the aim is to separate OOD|ID✗ID✓, whilst for SCOD the data is grouped as OODID✗|ID✓.

which is the average loss of the accepted samples. We are only concerned with the relative cost of ID✗ and OOD samples, so we use a single parameter β.

The objective is to find a classifier and rejection function (f, g) that minimise $R(f, g; t)$ for some given setting of t. We focus on comparing post-hoc (after-training) methods in this work, where g or equivalently S is varied with f fixed. This removes confounding factors that may arise from the interactions of different training-based and post-hoc methods, as they can often be freely combined.

In practice, both α and β will depend on the deployment scenario. However, whilst β can be set freely by the practitioner, α is outside of the practitioner's control and their knowledge of it is likely to be very limited.

It is worth contrasting the SCOD problem setting with OOD detection. SCOD aims to separate OOD, ID✗ |ID✓, whilst for OOD detection the data is grouped as OOD|ID✗, ID✓ (see Fig. 1). We note that previous work [26, 34, 35, 38, 41] refer to different types of predictive uncertainty, namely aleatoric and epistemic. The former arises from uncertainty inherent in the data (i.e. the true conditional distribution $P_{\text{tr}}(y|\boldsymbol{x})$) and as such is irreducible, whilst the latter can be reduced by having the model learn from additional data. Typically, it is argued that it is useful to distinguish these types of uncertainty at prediction time. For example, epistemic uncertainty should be an indicator of whether a test input $\boldsymbol{x}^*$ is OOD, whilst aleatoric uncertainty should reflect the level of class ambiguity of an ID input. An interesting result within our problem setting is that the conflation of these different types of uncertainties may not be an issue, as there is no need to separate ID✗ from OOD, as both should be rejected.

3 OOD Detectors Applied to SCOD

As the explicit objective of OOD detection is different to SCOD, it is of interest to understand how existing detection methods behave for SCOD. Previous work [27] has empirically shown that some existing OOD detection approaches perform worse, and in this section we shed additional light as to why this is the case.

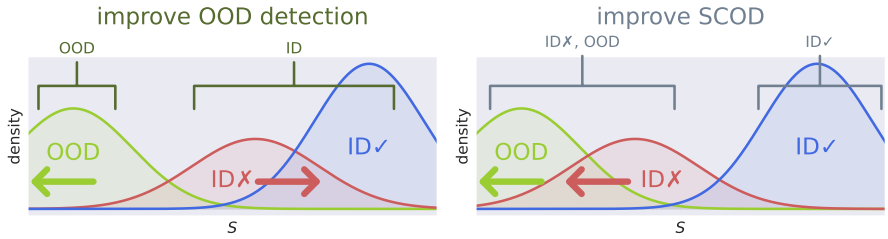

Fig. 2. Illustrations of how a detection method can improve over a baseline. Left: For OOD detection we can either have OOD further away from ID✓ or ID✗ closer to ID✓. Right: For SCOD we want both OOD and ID✗ to be further away from ID✓. Thus, we can see how improving OOD detection may in fact be at odds with SCOD.

Improving Performance: OOD Detection vs SCOD In order to build an intuition, we can consider, qualitatively, how detection methods can improve performance over a baseline, with respect to the distributions of OOD and ID✗ relative to ID✓. This is illustrated in Fig. 2. For OOD detection the objective is to better separate the distributions of ID and OOD data. Thus, we can either find a confidence score S that, compared to the baseline, has OOD distributed further away from ID✓, and/or has ID✗ distributed closer to ID✓. In comparison, for SCOD, we want both OOD and ID✗ to be distributed further away from ID✓ than the baseline. Thus there is a conflict between the two tasks as, for ID✗, the desired behaviour of confidence score S will be different.

Existing Approaches Sacrifice SCOD by Conflating ID✓ and ID✗ Considering post-hoc methods, the baseline confidence score S used is Maximum Softmax Probability (MSP) [16]. Improvements in OOD detection are often achieved by moving away from the softmax $\boldsymbol{\pi}$ in order to better capture the differences between ID and OOD data. Energy [33] and Max Logit [14] consider the logits $\boldsymbol{v}$ directly, whereas the Mahalanobis detector [31] and DDU [38] build generative models using Gaussians over the features $\boldsymbol{z}$. ViM [48] and Gradnorm [21] incorporate class-agnostic, feature-based information into their scores.

Recall that typically a neural network classifier learns a model $P(y|\boldsymbol{x};\boldsymbol{\theta})$ to approximate the true conditional distribution $P_{\text{tr}}(y|\boldsymbol{x})$ of the training data (Eqs. 2,3). As such, scores S extracted from the softmax outputs $\boldsymbol{\pi}$ should best reflect how likely a prediction on ID data is going to be correct or not (and this is indeed the case in our experiments in Section 5). As the above (post-hoc) OOD detection approaches all involve moving away from the modelled $P(y|\boldsymbol{x};\boldsymbol{\theta})$, we would expect worse separation between ID✗ and ID✓ even if overall OOD is better distinguished from ID. Fig. 3 shows empirically how well different types of data are separated using MSP ($\pi_{\max}$) and Energy ($\log \sum_k \exp v_k$), by plotting false positive rate (FPR) against true positive rate (TPR). Lower FPR indicates better separation of the negative class away from the positive class. Although

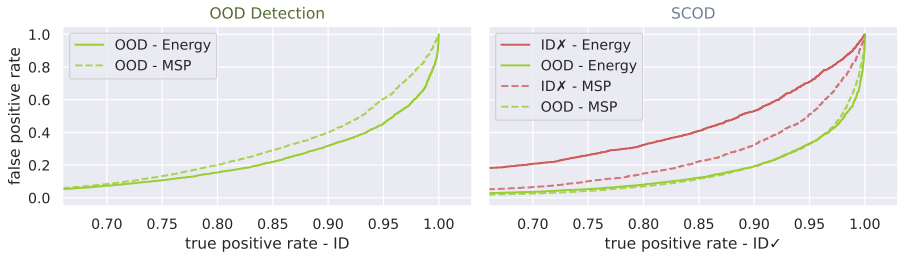

Fig. 3. Left: False positive rate (FPR) of OOD samples plotted against true positive rate (TPR) of ID samples. Energy performs better (lower) for OOD detection relative to the MSP baseline. Right: FPR of ID✗ and OOD samples against TPR of ID✓. Energy is worse than the baseline at separating ID✗|ID✓ and no better for OOD|ID✓, meaning it is worse for SCOD. Energy's improved OOD detection performance arises from pushing ID✗ closer to ID✓. The ID dataset is ImageNet-200, OOD dataset is iNaturalist and the model is ResNet-50.

Energy has better OOD detection performance compared to MSP, this is actually because the separation between ID✗ and ID✓ is much less for Energy, whilst the behaviour of OOD relative to ID✓ is not meaningfully different to the MSP baseline. Therefore, SCOD performance for Energy is worse in this case. Another way of looking at it would be that for OOD detection, MSP does worse as it conflates ID with OOD, however, this doesn't harm SCOD performance as much, as those ID samples are mostly incorrect anyway. The ID dataset is ImageNet-200 [27], OOD dataset is iNaturalist [22] and the model is ResNet-50 [13].

4 Targeting SCOD – Retaining Softmax Information

We would now like to develop an approach that is tailored to the task of SCOD. We have discussed how we expect softmax-based methods, such as MSP, to perform best for distinguishing ID✗ from ID✓, and how existing approaches for OOD detection improve over the baseline, in part, by sacrificing this. As such, to improve over the baseline for SCOD, we will aim to *retain* the ability to separate ID✗ from ID✓ whilst *increasing* the separation between OOD and ID✓.

Combining Confidence Scores Inspired by Gradnorm [21] and ViM [48] we consider the combination of two different confidence scores S_1, S_2. We shall consider S_1 our primary score, which we wish to augment by incorporating S_2. For S_1 we investigate scores that are strong for selective classification on ID data, but are also capable of detecting OOD data – MSP and (the negative of) softmax entropy, $(-)\mathcal{H}[\boldsymbol{\pi}]$. For S_2, the score should be useful *in addition* to S_1 in determining whether data is OOD or not. We should consider scores that capture different information about OOD data to the post-softmax S_1 if we want to improve OOD|ID✓. We choose to examine the l_1-norm of the feature vector $||\boldsymbol{z}||_1$

from [21] and the negative of the Residual[1] score $-||\boldsymbol{z}^{P^\perp}||_2$ from [48] as these scores capture class-agnostic information at the feature level. Note that although $||\boldsymbol{z}||_1$ and Residual have previously been shown to be useful for OOD detection in [21, 48], we do not expect them to be useful for identifying misclassifications. They are separate from the classification layer defined by $(\boldsymbol{W}, \boldsymbol{b})$, so they are far removed from the categorical $P(y|\boldsymbol{x};\boldsymbol{\theta})$ modelled by the softmax.

Softmax Information Retaining Combination (SIRC) We want to create a combined confidence score $C(S_1, S_2)$ that retains S_1's ability to distinguish ID✗ |ID✓ but is also able to incorporate S_2 in order to augment OOD|ID✓. We develop our approach based on the following set of *assumptions*:

- S_1 will be higher for ID✓ and lower for ID✗ and OOD.
- S_1 is bounded by maximum value $S_1^{\max}$. [2]
- S_2 is unable to distinguish ID✗ |ID✓, but is lower for OOD compared to ID.
- S_2 is useful in addition to S_1 for separating OOD|ID.

We propose to combine S_1 and S_2 using

$$C(S_1, S_2) = -(S_1^{\max} - S_1)\left(1 + \exp(-b[S_2 - a])\right) \ , \text{[3]} \tag{8}$$

where a, b are parameters chosen by the practitioner. The idea is for the accept/reject decision boundary of C to be in the shape of a sigmoid on the (S_1, S_2)-plane (See Fig. 4). As such the behaviour of only using the softmax-based S_1 is recovered for ID✗ |ID✓ as S_2 is increased, as the decision boundary tends to a vertical line. However, S_2 is considered increasingly important as it is decreased, allowing for improved OOD|ID✓. We term this approach Softmax Information Retaining Combination (SIRC).

The parameters a, b allow the method to be adjusted to different distributional properties of S_2. Rearranging Eq. 8,

$$S_1 = S_1^{\max} + C/[1 + \exp(-b[S_2 - a])] \ , \tag{9}$$

we see that a controls the vertical placement of the sigmoid, and b the sensitivity of the sigmoid to S_2. We use the empirical mean and standard deviation of S_2, μ_{S_2}, σ_{S_2} on ID data (training or validation) to set the parameters. We choose $a = \mu_{S_2} - 3\sigma_{S_2}$ so the centre of the sigmoid is below the ID distribution of S_2, and we set $b = 1/\sigma_{S_2}$, to match the ID variations of S_2. Note that other parameter settings are possible, and practitioners are free to tune a, b however they see fit (on ID data), but we find the above approach to be empirically effective.

Fig. 4 compares different methods of combination by plotting ID✓, ID✗ and OOD data densities on the (S_1, S_2)-plane. Other than SIRC we consider the

[1] $\boldsymbol{z}^{P^\perp}$ is the component of the feature vector that lies outside of a principle subspace calculated using ID data. For more details see Wang et al. [48]'s paper.

[2] This holds for our chosen S_1 of $\pi_{\max}$ and $-\mathcal{H}$.

[3] To avoid overflow this is implemented using the `logaddexp` function in PyTorch [40].

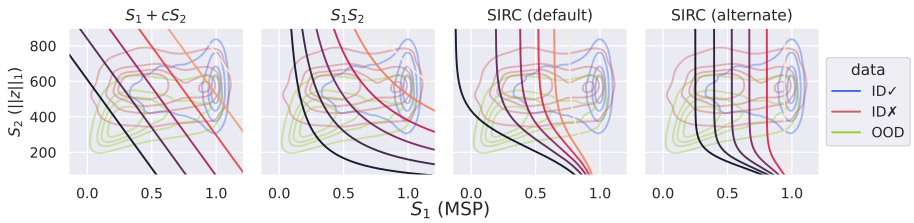

Fig. 4. Comparison of different methods of combining confidence scores S_1, S_2 for SCOD. OOD, ID✗ and ID✓ distributions are displayed using kernel density estimate contours. Graded contours for the different combination methods are then overlayed (lighter means higher combined score). We see that our method, SIRC (centre right) is able to better retain ID✗|ID✓ whilst improving OOD|ID✓. An alternate parameter setting for SIRC, with a stricter adherence to S_1, is also shown (far right). The ID dataset is ImageNet-200, the OOD dataset iNaturalist and the model ResNet-50. SIRC parameters are found using ID training data; the plotted distributions are test data.

combination methods used in ViM, $C = S_1 + cS_2$, where c is a user set parameter, and in Gradnorm, $C = S_1 S_2$. The overlayed contours of C represent decision boundaries for values of t. We see that the linear decision boundary of $C = S_1 + cS_2$ must trade-off significant performance in ID✗ |ID✓ in order to gain OOD|ID✓ (through varying c), whilst $C = S_1 S_2$ sacrifices the ability to separate ID✗ |ID✓ well for higher values of S_1. We also note that $C = S_1 S_2$ is not robust to different ID means of S_2. For example, arbitrarily adding a constant D to S_2 will completely change the behaviour of the combined score. On the other hand, SIRC is designed to be robust to this sort of variation between different S_2. Fig. 4 also shows an alternative parameter setting for SIRC, where a is lower and b is higher. Here more of the behaviour of only using S_1 is preserved, but S_2 contributes less. It is also empirically observable that the assumption that S_2 (in this case $||\boldsymbol{z}||_1$) is not useful for distinguishing ID✓ from ID✗ holds, and in practice this can be verified on ID validation data when selecting S_2.

We also note that although we have chosen specific S_1, S_2 in this work, SIRC can be applied to any S that satisfy the above assumptions. As such it has the potential to improve beyond the results we present, given better individual S.

5 Experimental Results

We present experiments across a range of CNN architectures and ImageNet-scale OOD datasets. Extended results can be found in the supplemental material.

Data, Models and Training For our ID dataset we use ImageNet-200 [27], which contains a subset of 200 ImageNet-1k [43] classes. It has separate training, validation and test sets. We use a variety of OOD datasets for our evaluation that display a wide range of semantics and difficulty in being identified. Near-ImageNet-200 (Near-IN-200) [27] is constructed from remaining ImageNet-1k

classes semantically similar to ImageNet-200, so it is especially challenging to detect. Caltech-45 [27] is a subset of the Caltech-256 [12] dataset with non-overlapping classes to ImageNet-200. Openimage-O [48] is a subset of the Open Images V3 [29] dataset selected to be OOD with respect to ImageNet-1k. iNaturalist [22] and Textures [48] are the same for their respective datasets [2, 47]. Colorectal [25] is a collection of histological images of human colorectal cancer, whilst Colonoscopy is a dataset of frames taken from colonoscopic video of gastrointestinal lesions [36]. Noise is a dataset of square images where the resolution, contrast and pixel values are randomly generated (for details see the supplemental material). Finally, ImageNet-O [18] is a dataset OOD to ImageNet-1k that is adversarially constructed using a trained ResNet. Note that we exclude a number of OOD datasets from [27] and [22] as a result of discovering ID examples.

We train ResNet-50 [13], DenseNet-121 [20] and MobileNetV2 [44] using hyperparameters based around standard ImageNet settings[4]. Full training details can be found in the supplemental material. For each architecture we train 5 models independently using random seeds $\{1, \ldots, 5\}$ and report the mean result over the runs. The supplemental material contains results on single pre-trained ImageNet-1k models, BiT ResNetV2-101 [28] and PyTorch DenseNet-121.

Detection Methods for SCOD We consider four variations of SIRC using the components $\{\text{MSP},\mathcal{H}\}\times\{||\boldsymbol{z}||_1,\text{Residual}\}$, as well as the components individually. We additionally evaluate various existing post-hoc methods: MSP [16], Energy [33], ViM [48] and Gradnorm [21]. For SIRC and ViM we use the full ID train set to determine parameters. Results for additional approaches, as well as further details pertaining to the methods, can be found in the supplemental material.

5.1 Evaluation Metrics

For evaluating different scoring functions S for the SCOD problem setting we consider a number of metrics. Arrows(↑↓) indicate whether higher/lower is better. (For illustrations and additional metrics see the supplemental material.)

Area Under the Risk-Recall curve (AURR)↓ We consider how empirical risk (Eq. 7) varies with recall of ID✓, and aggregate performance over different t by calculating the area under the curve. As recall is only measured over ID✓, the base accuracy of f is not properly taken into account. Thus, this metric is only suitable for comparing different g with f fixed. To give an illustrative example, a f, g pair where the classifier f is only able to produce a single correct prediction will have perfect AURR as long as S assigns that correct prediction the highest confidence (lowest uncertainty) score. Note that results for the AURC metric [10, 27] can be found in the supplemental material, although we omit them from the main paper as they are not notably different to AURR.

Risk@Recall=0.95 (Risk@95)↓ Since a rejection threshold t must be selected at deployment, we also consider a particular setting of t such that 95% of ID✓

[4] https://github.com/pytorch/examples/blob/main/imagenet/main.py

is recalled. In practice, the corresponding value of t could be found on a labelled ID validation set before deployment, without the use of any OOD data. It is worth noting that differences tend to be greater for this metric between different S as it operates around the tail of the positive class.

Area Under the ROC Curve (AUROC)↑ Since we are interested in rejecting both ID✗ and OOD, we can consider ID✓ as the positive class, and ID✗, OOD as separate negative classes. Then we can evaluate the AUROC of OOD|ID✓ and ID✗ |ID✓ independently. The AUROC for a specific value of α would then be a weighted average of the two different AUROCs. This is not a direct measure of risk, but does measure the separation between different empirical distributions. Note that due to similar reasons to AURR this method is only valid for fixed f.

False Positive Rate@Recall=0.95 (FPR@95)↓ FPR@0.95 is similar to AUROC, but is taken at a specific t. It measures the proportion of the negative class accepted when the recall of the positive class (or true positive rate) is 0.95.

5.2 Separation of ID✗ |ID✓ and OOD|ID✓ Independently

Table 1 shows %AUROC and %FPR@0.95 with ID✓ as the positive class and ID✗, OOD independently as different negative classes (see Section 5.1). In general, we see that SIRC, compared to S_1, is able to improve OOD|ID✓ whilst incurring only a small (< 0.2%AUROC) reduction in the ability to distinguish ID✗ |ID✓, across all 3 architectures. On the other hand, non-softmax methods designed for OOD detection show poor ability to identify ID✗, with performance ranging from ~ 8 worse %AUROC than MSP to ~ 50% AUROC (random guessing). Furthermore, they cannot consistently outperform the baseline when separating OOD|ID✓, in line with the discussion in Section 3.

SIRC is Robust to Weak S_2 Although for the majority of OOD datasets SIRC is able to outperform S_1, this is not always the case. For these latter instances, we can see that S_2 individually is not useful, e.g. for ResNet-50 on Colonoscopy, Residual performs *worse* than random guessing. However, in cases like this the performance is still close to that of S_1. As S_2 will tend to be higher for these OOD datasets, the behaviour is like that for ID✗ |ID, with the decision boundaries close to vertical (see Fig. 4). As such SIRC is *robust* to S_2 performing poorly, but is able to improve on S_1 when S_2 is of use. In comparison, ViM, which linearly combines Energy and Residual, is much more sensitive to when the latter stumbles. On Colonoscopy ViM has ~ 30 worse %FPR@95 compared to Energy, whereas SIRC ($-\mathcal{H}$, Res.) loses < 1% compared to $-\mathcal{H}$.

OOD Detection Methods are Inconsistent Over Different Data The performance of existing methods for OOD detection relative to the MSP baseline is varies considerably from dataset to dataset. For example, even though ViM is able to perform very well on Textures, Noise and ImageNet-O (>50 better %FPR@95 on Noise), it does worse than the baseline on most other OOD datasets (>20 worse %FPR@95 for Near-ImageNet-200 and iNaturalist). This

Table 1. %AUROC and %FPR@95 with ID✓ as the positive class, considering ID✗ and each OOD dataset separately. Full results are for ResNet-50 trained on ImageNet-200. We show abridged results for MobileNetV2 and DenseNet-121. **Bold** indicates best performance, underline 2nd or 3rd best and we show the mean over models from 5 independent training runs. Variants of SIRC are shown as tuples of their components (S_1,S_2). We also show error rate on ID data. SIRC is able to consistently match or improve over S_1 for OOD|ID✓, at a negligible cost to ID✗ |ID✓. Existing OOD detection methods are significantly worse for ID✗ |ID✓ and inconsistent at improving OOD|ID✓.

Model		Method	ID✗ AUROC↑	ID✗ FPR@95↓	OOD mean AUROC↑	OOD mean FPR@95↓	Near-IN-200 AUROC↑	Near-IN-200 FPR@95↓	Caltech-45 AUROC↑	Caltech-45 FPR@95↓	Openimage-O AUROC↑	Openimage-O FPR@95↓	iNaturalist AUROC↑	iNaturalist FPR@95↓
ResNet-50 ID %Error: 19.01	SIRC	(MSP,$\|\|\boldsymbol{z}\|\|_1$)	90.34	52.70	91.51	40.27	85.56	59.76	91.36	41.44	92.28	41.36	94.80	29.60
	SIRC	(MSP,Res.)	**90.43**	**52.10**	92.56	34.98	85.52	60.03	91.19	42.27	92.57	39.95	94.10	33.55
	SIRC	($-\mathcal{H}$,$\|\|\boldsymbol{z}\|\|_1$)	90.00	54.26	92.24	35.85	85.88	58.50	**92.19**	**36.08**	92.87	37.83	**95.38**	**25.09**
	SIRC	($-\mathcal{H}$,Res.)	90.13	54.01	**93.36**	**30.05**	85.85	58.93	92.11	36.76	**93.25**	**36.36**	94.82	28.51
		MSP	90.41	52.13	91.00	43.25	85.59	59.74	91.13	42.72	91.95	43.55	94.23	33.21
		$-\mathcal{H}$	90.07	54.05	91.81	38.24	**85.91**	**58.47**	92.01	37.20	92.59	40.10	94.90	28.01
		$\|\|\boldsymbol{z}\|\|_1$	48.06	94.70	78.22	58.70	52.27	94.58	70.28	77.83	72.23	71.51	85.65	49.50
		Residual	47.59	96.45	58.45	78.97	44.30	96.79	47.76	94.83	59.65	86.85	40.07	97.32
		Energy	82.05	69.79	92.06	35.32	81.96	68.70	92.15	38.62	90.92	46.28	94.13	31.70
		Gradnorm	60.17	87.88	85.22	44.41	62.90	86.89	81.11	59.23	81.09	57.80	91.00	34.46
		ViM	80.62	78.13	92.34	38.14	78.90	80.30	90.54	54.70	91.87	43.84	90.13	56.97

Model		Method	ID✗ AUROC↑	ID✗ FPR@95↓	Textures AUROC↑	Textures FPR@95↓	Colonoscopy AUROC↑	Colonoscopy FPR@95↓	Colorectal AUROC↑	Colorectal FPR@95↓	Noise AUROC↑	Noise FPR@95↓	ImageNet-O AUROC↑	ImageNet-O FPR@95↓
ResNet-50 ID %Error: 19.01	SIRC	(MSP,$\|\|\boldsymbol{z}\|\|_1$)	90.34	52.70	93.64	32.02	95.93	25.33	95.84	24.39	90.72	49.63	83.44	58.91
	SIRC	(MSP,Res.)	**90.43**	**52.10**	96.00	19.81	95.52	27.31	95.32	26.97	98.21	10.97	84.62	53.99
	SIRC	($-\mathcal{H}$,$\|\|\boldsymbol{z}\|\|_1$)	90.00	54.26	94.38	27.38	96.97	16.87	96.71	18.71	91.74	45.84	84.01	56.34
	SIRC	($-\mathcal{H}$,Res.)	90.13	54.01	96.68	15.70	96.72	18.10	96.41	20.42	99.02	4.89	85.33	50.81
		MSP	90.41	52.13	92.88	36.61	95.75	26.52	94.86	30.28	89.33	56.83	83.29	59.78
		$-\mathcal{H}$	90.07	54.05	93.77	30.79	96.87	17.55	95.93	23.43	90.47	51.63	83.89	57.02
		$\|\|\boldsymbol{z}\|\|_1$	48.06	94.70	88.90	39.67	76.97	82.24	97.28	14.64	97.36	13.51	63.00	84.82
		Residual	47.59	96.45	82.84	46.63	38.09	99.64	53.93	88.78	91.31	20.92	68.04	78.98
		Energy	82.05	69.79	95.37	22.50	**97.51**	**14.19**	**99.07**	5.00	94.93	29.05	82.52	61.86
		Gradnorm	60.17	87.88	93.00	26.57	90.54	42.85	98.98	**4.98**	97.59	13.05	70.78	73.88
		ViM	80.62	78.13	**98.46**	**7.62**	94.42	44.55	98.04	8.84	**99.82**	**0.31**	**88.85**	**46.15**

Model		Method	ID✗ AUROC↑	ID✗ FPR@95↓	OOD mean AUROC↑	OOD mean FPR@95↓
MobileNetV2 ID %Error: 21.35	SIRC	(MSP,$\|\|\boldsymbol{z}\|\|_1$)	89.53	55.51	92.27	34.82
	SIRC	(MSP,Res.)	**89.67**	55.10	91.78	38.56
	SIRC	($-\mathcal{H}$,$\|\|\boldsymbol{z}\|\|_1$)	88.90	58.64	**92.92**	**32.16**
	SIRC	($-\mathcal{H}$,Res.)	89.12	57.85	92.69	34.20
		MSP	89.64	**55.03**	91.54	39.73
		$-\mathcal{H}$	89.02	58.43	92.37	36.04
		$\|\|\boldsymbol{z}\|\|_1$	53.56	93.40	81.06	53.50
		Residual	41.99	97.30	41.42	94.11
		Energy	81.87	67.98	91.68	36.68
		Gradnorm	65.27	85.73	87.25	40.67
		ViM	80.21	74.36	89.46	51.97

Model		Method	ID✗ AUROC↑	ID✗ FPR@95↓	OOD mean AUROC↑	OOD mean FPR@95↓
DenseNet-121 ID %Error: 17.20	SIRC	(MSP,$\|\|\boldsymbol{z}\|\|_1$)	90.22	52.41	91.68	38.83
	SIRC	(MSP,Res.)	90.20	52.42	92.81	32.68
	SIRC	($-\mathcal{H}$,$\|\|\boldsymbol{z}\|\|_1$)	89.95	53.96	92.42	32.92
	SIRC	($-\mathcal{H}$,Res.)	89.92	54.17	**93.45**	**27.97**
		MSP	**90.30**	**51.85**	91.44	40.44
		$-\mathcal{H}$	90.04	53.41	92.24	34.49
		$\|\|\boldsymbol{z}\|\|_1$	36.87	98.70	63.53	80.35
		Residual	46.08	95.44	69.38	71.33
		Energy	82.12	66.54	90.92	38.87
		Gradnorm	50.18	95.19	76.18	62.58
		ViM	76.63	84.73	90.50	44.71

suggests that the inductive biases incorporated, and assumptions made, when designing existing OOD detection methods may prevent them from generalising across a wider variety of OOD data. In contrast, SIRC more *consistently*, albeit modestly, improves over the baseline, due to its aforementioned robustness.

5.3 Varying the Importance of OOD Data Through α and β

At deployment, there will be a specific ratio of ID:OOD data exposed to the model. Thus, it is of interest to investigate the risk over different values of α (Eq. 5). Similarly, an incorrect ID prediction may or may not be more costly than a prediction on OOD data so we investigate different values of β (Eq. 6). Fig. 5 shows how AURR and Risk@95 are affected as α and β are varied independently (with the other fixed to 0.5). We use the full test set of ImageNet-200, and pool

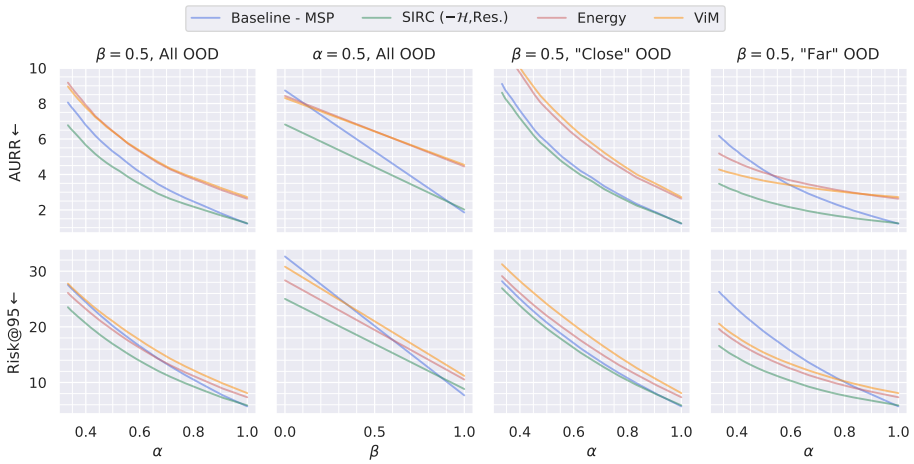

Fig. 5. AURR↓ and Risk@95↓ ($\times 10^2$) for different methods as α and β vary (Eqs. 5,6) on a mixture of all the OOD data. We also split the OOD data into qualitatively "Close" and "Far" subsets (Section 5.3). For high α, β, where ID✗ dominates in the risk, the MSP baseline is the best. As α, β decrease, increasing the effect of OOD data, other methods improve relative to the baseline. SIRC is able to *most consistently* improve over the baseline. OOD detection methods perform better on "Far" OOD. The ID dataset is ImageNet-200, the model ResNet-50. We show the mean over 5 independent training runs. We multiply all values by 10^2 for readability.

OOD datasets together and sample different quantities of data randomly in order to achieve different values of α. We use 3 different groupings of OOD data: All, "Close" {Near-ImageNet-200, Caltech-45, Openimage-O, iNaturalist} and "Far" {Textures, Colonoscopy, Colorectal, Noise}. These groupings are based on relative qualitative semantic difference to the ID dataset (see supplemental material for example images from each dataset). Although the grouping is not formal, it serves to illustrate OOD data-dependent differences in SCOD performance.

Relative Performance of Methods Changes with α and β At high α and β, where ID✗ dominates the risk, the MSP baseline performs best. However, as α and β are decreased, and OOD data is introduced, we see that other methods improve relative to the baseline. There may be a *crossover* after which the ability to better distinguish OOD|ID✓ allows a method to surpass the baseline. Thus, which method to choose for deployment will depend on the practitioner's setting of β and (if they have any knowledge of it at all) of α.

SIRC Most Consistently Improves Over the Baseline SIRC ($-\mathcal{H}$, Res.) is able to outperform the baseline most consistently over the different scenarios and settings of α, β, only doing worse for ID✗ dominated cases (α, β close to 1). This is because SIRC has close to baseline ID✗ |ID✓ performance and is superior

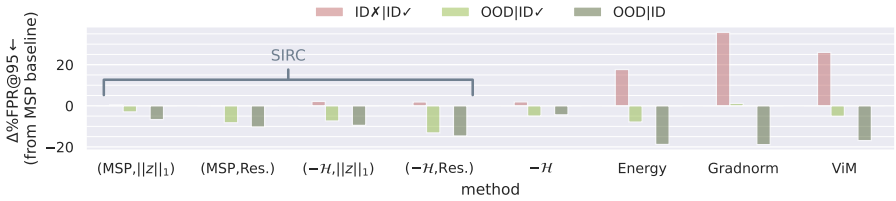

Fig. 6. The change in %FPR@95↓ relative to the MSP baseline of different methods. Different data classes are shown negative|positive. Although OOD detection methods are able to improve OOD|ID, they do so mainly at the expense of ID✗ |ID✓ rather than improving OOD|ID✓. SIRC is able to improve OOD|ID✓ with minimal loss to ID✗ |ID✓, alongside modest improvements for OOD|ID. Results for OOD are averaged over all OOD datasets. The ID dataset is ImageNet-200 and the model ResNet-50.

for OOD|ID✓. In comparison, ViM and Energy, which conflate ID✗ and ID✓, are often worse than the baseline for most (if not all) values of α, β. Their behaviour on the different groupings of data illustrates how these methods may be biased towards different OOD datasets, as they significantly outperform the baseline at lower α for the "Far" grouping, but always do worse on "Close" OOD data.

5.4 Comparison Between SCOD and OOD Detection

Fig. 6 shows the difference in %FPR@95 relative to the MSP baseline for different combinations of negative|positive data classes (ID✗ |ID✓, OOD|ID✓, OOD|ID), where OOD results are averaged over all datasets and training runs. In line with the discussion in Section 3, we observe that the non-softmax OOD detection methods are able to improve over the baseline for OOD|ID, but this comes mostly at the cost of inferior ID✗ |ID✓ rather than due to better OOD|ID✓, so they will do worse for SCOD. SIRC on the other hand is able to retain much more ID✗ |ID✓ performance whilst improving on OOD|ID✓, allowing it to have better OOD detection *and* SCOD performance compared to the baseline.

6 Related Work

There is extensive existing research into OOD detection, a survey of which can be found in [49]. To improve over the MSP baseline in [16], early post-hoc approaches, primarily experimenting on CIFAR-scale data, such as ODIN [32], Mahalanobis [31], Energy [33] explore how to extract non-softmax information from a trained network. More recent work has moved to larger-scale image datasets [14, 22]. Gradnorm [21], although motivated by the information in gradients, at its core combines information from the softmax and features together. Similarly, ViM [48] combines Energy with the class-agnostic Residual score. ReAct [45] aims to improve logit/softmax-based scores by clamping the magnitude of final

layer features. There are also many training-based approaches. Outlier Exposure [17] explores training networks to be uncertain on "known" existing OOD data, whilst VOS [4] instead generates virtual outliers during training for this purpose. [19, 46] propose the network explicitly learn a scaling factor for the logits to improve softmax behaviour. There also exists a line of research that explores the use of generative models, $p(\boldsymbol{x};\boldsymbol{\theta})$, for OOD detection [1, 39, 42, 50], however, these approaches are completely separate from classification.

Selective classification, or misclassification detection, has also been investigated for deep learning scenarios. Initially examined in [8, 16], there are a number of approaches to the task that target the classifier f through novel training losses and/or architectural adjustments [3, 9, 37]. Post-hoc approaches are fewer. DOCTOR [11] provides theoretical justification for using the l_2-norm of the softmax output $||\boldsymbol{\pi}||_2$ as a confidence score for detecting misclassifications, however, we find its behaviour similar to MSP and $\mathcal{H}$ (see supplemental material).

There also exist general approaches for uncertainty estimation that are then evaluated using the above tasks, e.g. Bayesian Neural Networks [23], MC-Dropout [7], Deep Ensembles [30], Dirichlet Networks [34, 35] and DDU [38].

The two works closest to ours are [24] and [27]. [24] investigates selective classification under covariate shift for the natural language processing task of question and answering. In the case of *covariate* shift, valid predictions can still be produced on the shifted data, which by our definition is not possible for OOD data (see Section 2). Thus the problem setting here is different to our work. We remark that it would be of interest to extend this work to investigate selective classification with covariate shift for tasks in computer vision. [27] introduces the idea that ID✗ and OOD data should be rejected together and investigates the performance of a range of existing approaches. They examine both training and post-hoc methods (comparing different f and g) on SCOD (which they term unknown detection), as well as misclassification detection and OOD detection. They do not provide a novel approach targeting SCOD, and consider a single setting of (α, β), where the α is not specified and $\beta = 0.5$.

7 Concluding Remarks

In this work, we consider the performance of existing methods for OOD detection on selective classification with out-of-distribution data (SCOD). We show how their improved OOD detection vs the MSP baseline often comes at the cost of inferior SCOD performance. Furthermore, we find their performance is inconsistent over different OOD datasets. In order to improve SCOD performance over the baseline, we develop SIRC. Our approach aims to retain information, which is useful for detecting misclassifications, from a softmax-based confidence score, whilst incorporating additional information useful for identifying OOD samples. Experiments show that SIRC consistently matches or improves over the baseline approach for a wide range of datasets, CNN architectures and problem scenarios.

Acknowledgements GX's PhD is funded jointly by Arm and the EPSRC.

References

[1] Caterini, A.L., Loaiza-Ganem, G.: Entropic issues in likelihood-based ood detection. ArXiv abs/2109.10794 (2021)

[2] Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., , Vedaldi, A.: Describing textures in the wild. In: Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2014)

[3] Corbière, C., THOME, N., Bar-Hen, A., Cord, M., Pérez, P.: Addressing failure prediction by learning model confidence. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32, pp. 2902–2913. Curran Associates, Inc. (2019), http://papers.nips.cc/paper/8556-addressing-failure-prediction-by-learning-model-confidence.pdf

[4] Du, X., Wang, Z., Cai, M., Li, Y.: Vos: Learning what you don't know by virtual outlier synthesis. ArXiv abs/2202.01197 (2022)

[5] El-Yaniv, R., Wiener, Y.: On the foundations of noise-free selective classification. J. Mach. Learn. Res. 11, 1605–1641 (2010)

[6] Fort, S., Ren, J., Lakshminarayanan, B.: Exploring the limits of out-of-distribution detection. In: NeurIPS (2021)

[7] Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: Balcan, M.F., Weinberger, K.Q. (eds.) Proceedings of The 33rd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 48, pp. 1050–1059. PMLR, New York, New York, USA (20–22 Jun 2016), https://proceedings.mlr.press/v48/gal16.html

[8] Geifman, Y., El-Yaniv, R.: Selective classification for deep neural networks. In: NIPS (2017)

[9] Geifman, Y., El-Yaniv, R.: Selectivenet: A deep neural network with an integrated reject option. In: International Conference on Machine Learning. pp. 2151–2159. PMLR (2019)

[10] Geifman, Y., Uziel, G., El-Yaniv, R.: Bias-reduced uncertainty estimation for deep neural classifiers. In: ICLR (2019)

[11] Granese, F., Romanelli, M., Gorla, D., Palamidessi, C., Piantanida, P.: Doctor: A simple method for detecting misclassification errors. In: NeurIPS (2021)

[12] Griffin, G., Holub, A., Perona, P.: Caltech-256 object category dataset (2007)

[13] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 770–778 (2016)

[14] Hendrycks, D., Basart, S., Mazeika, M., Mostajabi, M., Steinhardt, J., Song, D.X.: Scaling out-of-distribution detection for real-world settings. arXiv: Computer Vision and Pattern Recognition (2020)

[15] Hendrycks, D., Dietterich, T.G.: Benchmarking neural network robustness to common corruptions and perturbations. ArXiv abs/1903.12261 (2019)
[16] Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. ArXiv abs/1610.02136 (2017)
[17] Hendrycks, D., Mazeika, M., Dietterich, T.G.: Deep anomaly detection with outlier exposure. ArXiv abs/1812.04606 (2019)
[18] Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., Song, D.X.: Natural adversarial examples. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 15257–15266 (2021)
[19] Hsu, Y.C., Shen, Y., Jin, H., Kira, Z.: Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 10948–10957 (2020)
[20] Huang, G., Liu, Z., Weinberger, K.Q.: Densely connected convolutional networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 2261–2269 (2017)
[21] Huang, R., Geng, A., Li, Y.: On the importance of gradients for detecting distributional shifts in the wild. In: NeurIPS (2021)
[22] Huang, R., Li, Y.: Mos: Towards scaling out-of-distribution detection for large semantic space. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 8706–8715 (2021)
[23] Jospin, L.V., Laga, H., Boussaid, F., Buntine, W., Bennamoun, M.: Hands-on bayesian neural networks—a tutorial for deep learning users. IEEE Computational Intelligence Magazine 17(2), 29–48 (2022)
[24] Kamath, A., Jia, R., Liang, P.: Selective question answering under domain shift. In: ACL (2020)
[25] Kather, J.N., Weis, C.A., Bianconi, F., Melchers, S.M., Schad, L.R., Gaiser, T., Marx, A., Zöllner, F.G.: Multi-class texture analysis in colorectal cancer histology. Scientific Reports 6 (2016)
[26] Kendall, A., Gal, Y.: What uncertainties do we need in bayesian deep learning for computer vision? In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 5580–5590. NIPS'17, Curran Associates Inc., Red Hook, NY, USA (2017)
[27] Kim, J., Koo, J., Hwang, S.: A unified benchmark for the unknown detection capability of deep neural networks. ArXiv abs/2112.00337 (2021)
[28] Kolesnikov, A., Beyer, L., Zhai, X., Puigcerver, J., Yung, J., Gelly, S., Houlsby, N.: Big transfer (bit): General visual representation learning. In: ECCV (2020)
[29] Krasin, I., Duerig, T., Alldrin, N., Ferrari, V., Abu-El-Haija, S., Kuznetsova, A., Rom, H., Uijlings, J., Popov, S., Veit, A., Belongie, S., Gomes, V., Gupta, A., Sun, C., Chechik, G., Cai, D., Feng, Z., Narayanan, D., Murphy, K.: Openimages: A public dataset for large-scale multi-label and multi-class image classification. Dataset available from https://github.com/openimages (2017)
[30] Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: NIPS (2017)

[31] Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In: NeurIPS (2018)

[32] Liang, S., Li, Y., Srikant, R.: Enhancing the reliability of out-of-distribution image detection in neural networks. arXiv: Learning (2018)

[33] Liu, W., Wang, X., Owens, J.D., Li, Y.: Energy-based out-of-distribution detection. ArXiv abs/2010.03759 (2020)

[34] Malinin, A., Gales, M.J.F.: Predictive uncertainty estimation via prior networks. In: NeurIPS (2018)

[35] Malinin, A., Mlodozeniec, B., Gales, M.J.F.: Ensemble distribution distillation. ArXiv abs/1905.00076 (2020)

[36] Mesejo, P., Pizarro, D., Abergel, A., Rouquette, O.Y., Béorchia, S., Poincloux, L., Bartoli, A.: Computer-aided classification of gastrointestinal lesions in regular colonoscopy. IEEE transactions on medical imaging (2016)

[37] Moon, J., Kim, J., Shin, Y., Hwang, S.: Confidence-aware learning for deep neural networks. In: ICML (2020)

[38] Mukhoti, J., Kirsch, A., van Amersfoort, J.R., Torr, P.H.S., Gal, Y.: Deterministic neural networks with appropriate inductive biases capture epistemic and aleatoric uncertainty. ArXiv abs/2102.11582 (2021)

[39] Nalisnick, E.T., Matsukawa, A., Teh, Y.W., Görür, D., Lakshminarayanan, B.: Do deep generative models know what they don't know? ArXiv abs/1810.09136 (2019)

[40] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems 32, pp. 8024–8035. Curran Associates, Inc. (2019)

[41] Pearce, T., Brintrup, A., Zhu, J.: Understanding softmax confidence and uncertainty. ArXiv abs/2106.04972 (2021)

[42] Ren, J., Liu, P.J., Fertig, E., Snoek, J., Poplin, R., Depristo, M., Dillon, J., Lakshminarayanan, B.: Likelihood ratios for out-of-distribution detection. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019), https://proceedings.neurips.cc/paper/2019/file/1e79596878b2320cac26dd792a6c51c9-Paper.pdf

[43] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.S., Berg, A.C., Fei-Fei, L.: Imagenet large scale visual recognition challenge. International Journal of Computer Vision 115, 211–252 (2015)

[44] Sandler, M., Howard, A.G., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 4510–4520 (2018)

[45] Sun, Y., Guo, C., Li, Y.: React: Out-of-distribution detection with rectified activations. In: NeurIPS (2021)
[46] Techapanurak, E., Suganuma, M., Okatani, T.: Hyperparameter-free out-of-distribution detection using cosine similarity. In: Proceedings of the Asian Conference on Computer Vision (ACCV) (November 2020)
[47] Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S.: The inaturalist species classification and detection dataset (2017), https://arxiv.org/abs/1707.06642
[48] Wang, H., Li, Z., Feng, L., Zhang, W.: Vim: Out-of-distribution with virtual-logit matching. ArXiv abs/2203.10807 (2022)
[49] Yang, J., Zhou, K., Li, Y., Liu, Z.: Generalized out-of-distribution detection: A survey. ArXiv abs/2110.11334 (2021)
[50] Zhang, M., Zhang, A., McDonagh, S.G.: On the out-of-distribution generalization of probabilistic image modelling. In: NeurIPS (2021)