Sample-free white-box out-of-distribution detection for deep learning Supplementary

Jean-Michel Begon Pierre Geurts
University of Liege
{jm.begon,p.geurts}@uliege.be

A. Considerations regarding the indicators

A.1. General remarks on optimality-based indicators

The first order optimality condition is such that (disregarding regularization for now)

$$\mathbb{E}_{x,y\sim\mathcal{I}}\nabla_\Theta E(x,y;\Theta) = 0 \tag{1}$$

Owing to the backpropagation, we have

$$\frac{\partial E(x,y;\Theta)}{\partial z_j} = p_j(x) - y_j \tag{2}$$

$$\nabla_{\theta_l} E(x,y;\Theta) = \left(\prod_{i=l}^{L} J_i^T(x)\right) \nabla_z E(x,y;\Theta) \tag{3}$$

where $J_i(x)$ is the Jacobian of the ith layer at x.

A gradient-based optimization of the network will not only push p_j toward y_j but will also tend to produce uniform probabilities where $y_j = 0$, since the parameters relating to higher gaps will be affected more.

If the network performs well, all the first order information should be contained in $\nabla_z E(x,y)$, as a Jacobian being responsible for zeroing the gradient would (i) not force optimality for layers further in the network, and (ii) would suggest a vanishing gradient situation leading to being trapped in a local minimum.

Since the network is optimized on a learning sample (and furthermore there is a regularization term), the expected gradient of the loss is not zero, in practice. However, a deployed network should have a small expected loss gradient, which allow us to derive several indicators for which we expect the values on ID samples to be low.

Optimality vs. performance Even though the indicators are qualified as optimality-based, they also rely on the network performing well, that is, we expect both the loss and its gradient to be small on ID samples. We feel this should be the case for any deployed model, especially in sensitive applications. In the event where the gradient of the loss is small while the loss itself is not so small (for instance, with a model of insufficient capacity), we expect the so-called "optimality-based" indicators to underperform. We discuss this further in Appendix C.1 when comparing CIFAR 10 and CIFAR 100. We additionnaly shed some more light on the matter in Appendix C.3, where we discuss the impact of the model quality, and in Appendix C.4, where we discuss the related problem of misclassification detection.

proj The PROJ indicator combines the information from both ANG and NORM. On datasets where there is no asymmetry (such as class imbalance, for instance), $||w_j||$ and b_j tend to be relatively constant with respect to j (see Table 1). Therefore PROJ is expected to be closely related to the logit. This is confirmed by Table 3 of the main paper.

positivity ReLU-based architectures, that is, most modern ones in image classification, end the feature extraction phase with a ReLU activation, possibly followed by max or average pooling. As a result, the latent vectors are non-negative, whereas some (significant part of the) components of the hyperplane weights are negative (see Table 1). As a consequence, most weights are used to bid *against* the other classes, rather than *for* the predicted one. This suggests that it might be worth looking at the positive and negative parts of the previous indicators separately.

Let $\mathcal{P}(w) = \{1 \leq i \leq p_{L-1} | w^{(i)} \geq 0\}$ be the set of indices of the non-negative components of w. Suppose the predicted class for x is k and let

Table 1: Statistics of the latent space parameters.

	order of std(W) / std(b)		Percentage of positive components of W	
	CIFAR 10	CIFAR 100	CIFAR 10	CIFAR 100
ResNet 50	10^{-2} / 10^{-1}	10^{-2} / 10^{-2}	37.5	35.8
WideResNet	10^{-1} / 10^{-1}	10^{-1} / 10^{-2}	41.1	40.1
DenseNet 121	10^{-1} / 10^{-1}	10^{-2} / 10^{-2}	42.1	43.1

$$||\cdot||_j^+ = \sqrt{\sum_{i\in\mathcal{P}(w_j)} \left(\cdot^{(i)}\right)^2} \tag{4}$$

$$\text{ACT+}(x) = -\sum_{i\in\mathcal{P}(k)} w_k^{(i)} u^{(i)} \tag{5}$$

$$\text{ANG++}(x) = 1 + \frac{\text{ACT+}(x)}{||w_k||_k^+ \, ||u||_k^+} \tag{6}$$

The rationale for using the positive indicators ACT+ and ANG++, instead of their counterpart, is to reject OOD samples whose high probability would be due to being unlikely to come from any other classes than the predicted one, rather than appearing to belong to the predicted class. Note that positivity also implies ANG cannot be zero.

A.2. Relationship between logit and T1000

The main result covered by this section is

$$p_{k|T=1000}(x) \approx \frac{c}{K} + \frac{1}{TK} z_k(x) \tag{7}$$

where k is the predicted class by the network, K is the number of class, $z(x)$ is the logit vector corresponding to input x and $T = 1000$ is the temperature. It holds so long as $||z|| \ll T$ and the network is trained long enough.

For shorthand, let $\hat{p}_k$ stands for softmax_k. From Taylor decomposition, it follows that

$$p_{k|T=1000} = \hat{p}_k\left(\frac{1}{T}z\right) \tag{8}$$

$$= \hat{p}_k(0) + \frac{1}{T}\left(\nabla_z \hat{p}_k\,(0)\right)^T z + o\left(\frac{1}{T^2}||z||\right) \tag{9}$$

$$\approx \frac{1}{K} + \frac{1}{T}\sum_{j=1}^{K}\left(\hat{p}_k(0)\left(\delta_{j,k} - \hat{p}_j(0)\right) z_j\right) \tag{10}$$

$$= \frac{1}{K} + \frac{1}{T}\sum_{j=1}^{K}\left(\frac{1}{K}\left(\delta_{j,k} - \frac{1}{K}\right) z_j\right) \tag{11}$$

$$= \frac{1}{K} + \frac{1}{T\,K}\left(z_k - \frac{1}{K}\sum_{j=1}^{K} z_j\right) \tag{12}$$

$$= \frac{1}{K} + \frac{z_k - \bar{z}}{T\,K} \tag{13}$$

where $\delta_{j,k}$ is the Kronecker symbol.

By linearity, we have

$$\bar{z}(x) = \frac{1}{K}\sum_j z_j(x) \tag{14}$$

$$= \frac{1}{K}\sum_j \left(w_j^T x + b_j\right) \tag{15}$$

$$= \overline{w}^T x + \bar{b} \tag{16}$$

$$\Delta z(x) = z_k(x) - \bar{z}(x) \tag{17}$$

$$= \left(w_k^T x + b_k\right) - \left(\overline{w}^T x + \bar{b}\right) \tag{18}$$

$$= \Delta w_k^T\, x + \Delta b_k \tag{19}$$

So we end up with a linear relationship between $p_{k|T=1000}$ and the Δ-logit Δz.

In addition, the average weight vector and bias tends to be close to null. Owing to the softmax translation invariance

$$\frac{e^{z_k - (\overline{w}^T x + \hat{b})}}{\sum_j e^{z_j - (\overline{w}^T x + \hat{b})}} = \frac{e^{z_k}}{\sum_j e^{z_j}} \tag{20}$$

the only incentive acting on the average weight vector and bias is the slight penalization which goes in the direction of $\overline{w} = 0$ and $\bar{b} = 0$. Indeed,

$$\overline{w} = \frac{1}{K}\left(w_k + \sum_{j\neq k} w_j\right) = \frac{1}{K}\left(w_k + w_{\neg k}\right) \tag{21}$$

Since $w_{\neg k}$ represents a hyperplane for rejecting class k, it would be wasteful for the network not enforcing $w_k = -w_{\neg k}$. As for $\bar{b}$, it does not depend on x and can be hidden away in the independent term.

Overall—provided the network was trained enough—we arrive at the conclusion

$$\overline{w}^T x \ll w_k^T x \tag{22}$$

$$\implies p_{k|T=1000} \approx \frac{c}{K} + \frac{z_k}{T\,K} \tag{23}$$

For the purpose of ranking samples, we can further remove the constant term $\overline{\mathcal{Z}} = \bar{z} - \epsilon(z)$, where $\overline{\mathcal{Z}}$ is the expected average of logits over the input space and $\epsilon(z)$ is the deviation of the logit mean from its expectation.

This leads to

$$= \frac{1}{K} + \frac{z_k - \left(\epsilon(z) + \overline{\mathcal{Z}}\right)}{T\,K} \tag{24}$$

$$= \frac{1}{K}\left(1 - \frac{\overline{\mathcal{Z}}}{T}\right) + \frac{z_k - \epsilon(z)}{T\,K} \tag{25}$$

$$= \frac{c'}{K} + \frac{z_k}{T\,K} - \frac{\epsilon(z)}{T\,K} \tag{26}$$

In this last relationship, $\epsilon(z)$ is of the order of magnitude of the standard deviation of $\overline{w}$ and $\overline{b}$, typically 8 order smaller than z_k (see Table 2) and can be safely ignored.

B. Protocol details

Here we further discuss the protocol we used. We do not expect our main results to be influenced much by the actual details, so long as the networks are well-trained.

In our empirical study, we relied on three networks and three image classification tasks to serve as ID datasets, namely CIFAR 10, CIFAR 100 [6] and ImageNet [3].

CIFARs CIFAR 10/CIFAR 100 consist in 60000 32×32 RGB images. CIFAR 10 has 10 classes, while CIFAR 100 has 100. The datasets are balanced class-wise (across both train and test sets). There is a standard train/test split of respectively 50000 and 10000 images. We split further the training set to keep 5000 images as validation set. The testing samples are only used to assess the model accuracies (Table 3) at the end of training; they are not used to back up any decision during training. They serve as ID samples for our experiments.

ImageNet In the case of ImageNet, We re-used the available weights through PyTorch [7]. The 100000 unlabeled RGB test images, spread among 1000 classes, are used as ID samples. We followed the standard procedure to rescale the images to a size of 256 along its shortest spatial dimension and extract centred 224×224 crops.

Networks The networks are a ResNet 50 [4], a WideResNet-40 [8] and a DenseNet 121 [5]. All three architectures are ReLU-based and output non-negative latent vectors. Table 3 portrays the accuracy of each model.

On CIFARs, they expect input of size 32×32 and were trained for 450 epochs by stochastic gradient descent (batches of size 128, weight decay of 5×10^{-4} and momentum of 0.9). The learning rate was initialized at 0.1. It was decreased by a factor 10 after 150 epochs and again at epoch 300. Each decrease was accompanied by a restart from the best model according to the validation accuracy. Horizontal flip and random cropping (with a padding of 4) were used as data augmentation.

As mentioned, PyTorch's pre-trained network were used for ImageNet and expected images of size 224×224.

Pre-processing Prior to running through the network, all images (ID and OOD) are resized to fit the network expected size, transformed to RGB if necessary, rescaled in the range $[0, 1]$ and then normalized channel-wise according to the ID dataset input statistics (see next paragraph). Artifacts due to resizing may help detect OOD samples. In the case of artificial datasets, images are generated with the appropriate size.

Input normalization For CIFARs, we estimated the channel mean/standard deviation on the training set. Regarding the standard deviation, we computed the square root of the *total* variance, in accordance to PyTorch's batchnorm implementation. For some reasons, available statistics usually used the average *intra-image* variance, disregarding the *inter-image* variance. Admittedly, the difference is slight.

For ImageNet, we re-used pre-trained network and thus conformed to using the same statistics as were used for training (based on intra-image variance).

ID/OOD balance Except in the case of the supervised approach (see main text), the whole ID test set and the whole OOD dataset are used to assess the indicator performances. As a consequence, the classification task is quite unbalanced (as might be the case in a real setting, although we might expect a much higher proprotion of ID samples). For artificial datasets, we generated 50000 samples.

Variability On CIFARs, results are established on three random initializations of the network's parameters and is, with batch sampling, the only sources of randomness; artificially-generated datasets are the same throughout the experiments. Since we re-used pre-trained models for ImageNet, there is only one experiment per network (there is only one set of weights available per architecture). Note that the IN- indicators are independent of the network; they only depend on the input, channel-wise statistics of the ID datasets. As such, they are not subject to randomness.

C. Additional results

Table 2: Mean values for the average weight vector and bias, as well as the logit of the predicted class (on the ID task). The first two tends to be very small, while the last one is several order of magnitude higher. Although the logit is expected to be lower on a OOD task, orders of magnitude are equivalent. C. 10 and C. 100 stand for CIFAR 10 and CIFAR 100 respectively.

		$\|\|\overline{w}\|\|$	$\bar{b}$	z_k
C. 10	ResNet 50	$5\ 10^{-6} \pm 4\ 10^{-7}$	$1\ 10^{-7} \pm 3\ 10^{-7}$	11.3 ± 2.5
	WideResNet	$4\ 10^{-6} \pm 1\ 10^{-6}$	$-3\ 10^{-7} \pm 3\ 10^{-7}$	12.0 ± 3.4
	DenseNet 121	$3\ 10^{-6} \pm 1\ 10^{-7}$	$-2\ 10^{-7} \pm 2\ 10^{-7}$	10.2 ± 2.2
C. 100	ResNet 50	$2\ 10^{-6} \pm 6\ 10^{-8}$	$1\ 10^{-8} \pm 5\ 10^{-9}$	13.3 ± 3.8
	WideResNet	$4\ 10^{-6} \pm 4\ 10^{-6}$	$3\ 10^{-7} \pm 2\ 10^{-7}$	13.5 ± 4.4
	DenseNet 121	$7\ 10^{-7} \pm 2\ 10^{-7}$	$5\ 10^{-8} \pm 4\ 10^{-8}$	13.2 ± 4.1

Table 3: Model Performance (in %) on the ID task.

	Accuracy		ImageNet	
	CIFAR 10	CIFAR 100	Top-1 error	Top-5 error
ResNet 50	94.11 ± 0.25	77.48 ± 0.23	23.85	7.13
WideResNet	94.18 ± 0.31	74.17 ± 0.72	21.49	5.91
DenseNet 121	94.30 ± 0.31	77.89 ± 0.04	25.35	7.83

C.1. Detailed auroc tables

Tables 4-6 holds detailed results for CIFAR 10, CIFAR 100 and Imagenet as ID datasets, respectively.

Supervised approach Although not the focus of this work, we see that a supervised linear SVM [2] established on the true ID/OOD mix distribution performs almost perfectly. ON CIFARs, it only struggles with Tiny ImageNet (TIN) and LSUN, where it still performs best, except on TIN with CIFAR 100. In that setting, the mean results of ACT, ANG and sometimes T1000 is slightly higher.

On ImageNet, it is perfect but for LSUN, where it comes first with a large margin.

Baselines ODIN and T1000 are strong baselines. For ImageNet, it would seem the additional perturbation provided by ODIN pays off, especially on grey images (fashion MNIST, MNIST). On CIFARs, the gap is much less present, apart on MNIST for CIFAR 100 where 5 to 10 percent of auroc are lost. On harder tasks, T1000 may have a slight edge.

MP and H rarely yield remarkable results.

Batchnorm indicators The IN- family of indicators may work well at detecting grey images, although IN-DSS never really works. When input statistics are close to the ID's (Tiny ImageNet, LSUN), those indicators do not work better than random. They also fail on the noisy Gaussian dataset, which has individual pixel statistics that are close to ID's. It would be easy to reject such samples if inter-channel information were available.

On the other hand, indicators based on all batchnorm layers work extremely well on Gaussian since intermediate tensors contain inter-channel information, thanks to convolutions. Interestingly, DSS and/or DSS-EXT perform well on SVHN in all settings. Those indicators are much less robust to the network and its initialization, however. For instance, DMS achieves $82.73 \pm 0.56\%$ on DenseNet 121 for discriminating fashion MNIST against CIFAR 10, but it only achieves $69.39 \pm 6.49\%$ on ResNet 50 for the same task (note the high variance).

Quite often, batchnorm indicators have auroc much lower than 50%, indicating lower values for OOD samples. In our sample-free setting, we can only discard such indicators and conclude they can only discriminate specific OOD sets. However, in a supervised setting, such indicators might prove useful as the ordering condition we impose on indicators could be altogether ignored.

Latent space indicators As expected, NORM and NORM+ do not convey the appropriate information. The remaining indicators rank well, however. On ImageNet, positive-only indicators seem to work better, while this is not as clear for the other ID tasks. In particular, ANG++ performs better than ANG on ImageNet but ANG works better in the other settings (except for ResNet 50 on CIFAR 100). Once again, the OOD dataset has an impact on the ranking: ACT/ACT+ tend to struggle with (fashion) MNIST on CIFAR 100 and ImageNet, while, with ImageNet as ID task, ANG++

comes way ahead of the other indicators against CIFARs as OOD but underperfoms on LSUN (except on WideResNet).

1C-Sum Overall, 1C-Sum results are good. Compared to individual indicators, it mainly lags behind on MNIST with CIFARs as ID sets and on LSUN. Hopefully, as soon as data become available, 1C-Sum can be turned into the supervised indicator to compensate for its initial weaknesses. More precisely, incorporating the IN- feature to better detect grey images and drop other batchnorm indicators for LSUN. Assumptions regarding the expected OOD distribution may also help tuning the model weights.

CIFAR 10 vs. CIFAR 100 CIFAR 100 is a harder base task than CIFAR 10 and even well-optimized networks achieve more modest performances (Table 3). As we can see, the gap in accuracy is reflected in OOD detection as well, at least on optimality-based indicators, thus confirming that we also need the loss to be small for OOD detection. In Appendix C.4, we will investigate whether the lower auroc scores can be attributed to lower accuracies.

Table 4: Area under the ROC curve for OOD detection with CIFAR 10 as ID. TIN stands for Tiny ImageNet.

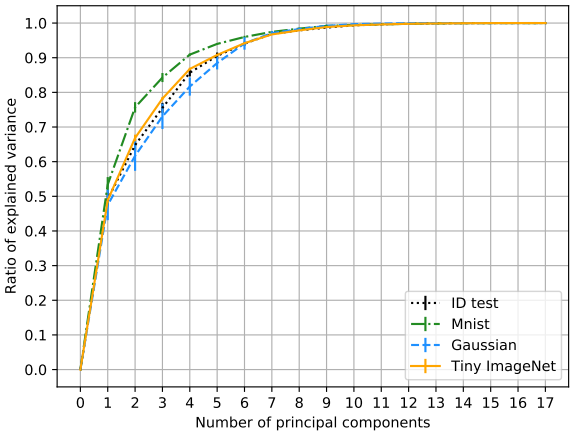
		Gaussian	SVHN	MNIST	fash. MNIST	TIN	LSUN
ResNet 50	ODIN	91.36 ± 5.42	90.22 ± 4.03	96.88 ± 0.70	95.89 ± 0.75	87.22 ± 2.12	92.38 ± 1.56
	T1000	83.17 ± 9.00	93.14 ± 3.05	94.81 ± 0.78	95.43 ± 0.62	88.70 ± 1.23	92.66 ± 1.04
	MP	89.27 ± 4.90	91.89 ± 1.30	90.76 ± 0.65	91.97 ± 0.47	87.05 ± 0.61	90.08 ± 0.60
	H	89.05 ± 5.03	92.51 ± 1.46	91.40 ± 0.62	92.71 ± 0.58	87.52 ± 0.67	90.62 ± 0.59
	NORM	53.96 ± 33.02	85.46 ± 10.89	92.28 ± 4.92	89.52 ± 4.00	80.19 ± 4.27	82.50 ± 4.93
	NORM+	54.99 ± 28.60	87.17 ± 9.12	94.61 ± 2.09	92.92 ± 1.85	85.00 ± 2.61	88.87 ± 2.82
	ACT	83.34 ± 9.02	93.32 ± 2.95	94.90 ± 0.70	95.47 ± 0.59	88.77 ± 1.18	92.50 ± 1.08
	ACT+	87.68 ± 9.18	94.23 ± 3.50	96.03 ± 1.44	95.93 ± 0.72	88.05 ± 1.53	91.68 ± 1.38
	PROJ	85.53 ± 8.09	94.01 ± 2.42	95.61 ± 0.40	95.47 ± 0.58	88.61 ± 1.26	92.05 ± 1.21
	ANG	91.78 ± 2.79	93.41 ± 0.09	94.15 ± 0.60	94.76 ± 1.02	88.35 ± 0.51	91.98 ± 0.58
	ANG++	99.89 ± 0.12	97.26 ± 0.17	94.25 ± 1.22	93.41 ± 1.70	86.05 ± 0.88	88.43 ± 0.75
	IN-DMS	7.85	60.46	98.59	71.94	52.89	49.26
	IN-DMS-AOS	52.79	30.41	99.68	96.02	52.55	54.91
	IN-DSS	5.13	85.99	36.16	58.53	52.03	42.94
	DMS	100.00 ± 0.00	80.29 ± 8.30	93.97 ± 2.47	69.39 ± 6.49	34.21 ± 5.54	22.67 ± 5.33
	DMS-AOS	99.25 ± 0.48	4.72 ± 2.26	81.12 ± 9.04	59.42 ± 9.53	25.25 ± 2.65	23.78 ± 2.66
	DSS	99.86 ± 0.14	96.51 ± 0.60	70.33 ± 15.30	62.22 ± 3.53	55.01 ± 1.90	47.40 ± 4.40
	DSS-EXT	98.24 ± 0.61	97.70 ± 0.34	66.93 ± 1.88	67.64 ± 1.67	66.84 ± 0.88	62.94 ± 1.38
	supervised	100.00 ± 0.00	99.75 ± 0.05	100.00 ± 0.00	99.70 ± 0.03	90.82 ± 0.45	96.14 ± 0.19
	1C-Sum	97.84 ± 2.70	97.83 ± 0.95	96.47 ± 1.58	95.86 ± 0.63	88.86 ± 0.79	91.61 ± 0.90
WideResNet	ODIN	99.73 ± 0.20	90.85 ± 5.11	94.11 ± 4.14	95.22 ± 1.16	84.31 ± 4.70	90.22 ± 2.87
	T1000	98.13 ± 1.37	95.20 ± 1.76	91.59 ± 4.71	94.88 ± 0.79	87.65 ± 2.06	91.49 ± 1.49
	MP	96.69 ± 1.60	93.34 ± 1.01	88.42 ± 3.64	92.01 ± 0.33	86.62 ± 1.04	89.69 ± 0.75
	H	97.41 ± 1.65	94.10 ± 1.25	89.08 ± 3.89	92.76 ± 0.38	87.09 ± 1.16	90.22 ± 0.85
	NORM	98.11 ± 2.24	92.21 ± 4.95	89.09 ± 10.60	87.96 ± 5.56	76.42 ± 7.85	79.48 ± 6.15
	NORM+	98.97 ± 0.90	93.73 ± 3.41	90.84 ± 7.45	92.89 ± 2.58	83.54 ± 4.64	87.58 ± 2.34
	ACT	98.32 ± 1.24	95.35 ± 1.71	91.96 ± 4.39	94.96 ± 0.77	87.73 ± 2.03	91.36 ± 1.50
	ACT+	99.17 ± 0.75	95.54 ± 2.36	92.45 ± 5.73	94.87 ± 1.46	85.57 ± 3.49	89.70 ± 2.72
	PROJ	98.29 ± 1.38	95.67 ± 1.46	92.88 ± 3.68	94.94 ± 0.69	87.74 ± 1.92	90.97 ± 1.64
	ANG	96.24 ± 1.90	92.63 ± 0.75	90.78 ± 1.20	93.49 ± 0.78	88.68 ± 0.50	91.61 ± 1.07
	ANG++	97.41 ± 2.15	92.46 ± 1.83	87.46 ± 3.74	86.36 ± 3.87	80.42 ± 0.15	81.92 ± 4.51
	IN-DMS	7.85	60.46	98.59	71.94	52.89	49.26
	IN-DMS-AOS	52.79	30.41	99.68	96.02	52.55	54.91
	IN-DSS	5.13	85.99	36.16	58.53	52.03	42.94
	DMS	100.00 ± 0.00	94.13 ± 0.87	98.26 ± 1.11	80.14 ± 3.04	48.94 ± 0.77	38.39 ± 0.75
	DMS-AOS	100.00 ± 0.00	4.10 ± 0.62	78.41 ± 2.76	53.01 ± 2.33	35.58 ± 1.63	40.37 ± 2.01
	DSS	100.00 ± 0.00	96.83 ± 0.27	83.13 ± 2.05	80.13 ± 2.09	54.11 ± 3.19	40.92 ± 4.76
	DSS-EXT	94.81 ± 1.16	97.77 ± 0.24	68.79 ± 2.13	71.78 ± 1.34	63.17 ± 2.63	53.89 ± 3.84
	supervised	100.00 ± 0.00	99.74 ± 0.04	100.00 ± 0.00	99.69 ± 0.02	90.64 ± 0.38	95.13 ± 0.64
	1C-Sum	100.00 ± 0.00	98.87 ± 0.19	94.98 ± 2.17	95.50 ± 0.73	87.49 ± 2.51	89.28 ± 2.89
DenseNet 121	ODIN	99.49 ± 0.37	82.99 ± 3.14	85.32 ± 8.29	88.69 ± 3.87	75.92 ± 1.61	82.35 ± 3.57
	T1000	96.93 ± 1.89	93.66 ± 1.47	84.19 ± 6.56	91.87 ± 1.89	83.51 ± 0.49	87.53 ± 2.03
	MP	96.49 ± 0.91	93.07 ± 1.30	85.76 ± 3.96	91.67 ± 0.71	85.48 ± 0.30	88.45 ± 1.00
	H	96.79 ± 1.22	93.62 ± 1.37	86.04 ± 4.10	92.14 ± 0.78	85.72 ± 0.35	88.75 ± 1.10
	NORM	51.24 ± 34.67	65.16 ± 6.79	44.44 ± 22.17	51.29 ± 15.42	46.78 ± 2.38	49.02 ± 4.55
	NORM+	65.46 ± 29.49	78.63 ± 4.36	59.79 ± 17.81	73.50 ± 11.20	66.09 ± 1.53	70.25 ± 3.43
	ACT	97.13 ± 1.78	93.86 ± 1.43	84.58 ± 6.46	91.94 ± 1.83	83.59 ± 0.49	87.32 ± 2.04
	ACT+	96.30 ± 3.39	90.53 ± 2.00	76.80 ± 11.96	85.04 ± 6.23	76.24 ± 0.83	80.38 ± 2.96
	PROJ	97.45 ± 1.61	94.39 ± 1.44	86.86 ± 5.04	92.07 ± 1.67	83.55 ± 0.42	86.22 ± 2.07
	ANG	98.75 ± 0.39	96.46 ± 0.79	94.65 ± 0.44	95.44 ± 0.88	89.68 ± 0.39	91.86 ± 0.71
	ANG++	99.85 ± 0.15	94.74 ± 1.24	95.67 ± 0.63	90.89 ± 1.15	83.01 ± 2.71	84.29 ± 2.09
	IN-DMS	7.85	60.46	98.59	71.94	52.89	49.26
	IN-DMS-AOS	52.79	30.41	99.68	96.02	52.55	54.91
	IN-DSS	5.13	85.99	36.16	58.53	52.03	42.94
	DMS	99.99 ± 0.00	94.61 ± 0.61	98.29 ± 0.36	82.73 ± 0.56	42.67 ± 1.58	34.95 ± 1.30
	DMS-AOS	98.79 ± 0.39	12.33 ± 1.91	74.75 ± 0.64	55.15 ± 1.00	27.51 ± 0.77	35.15 ± 0.31
	DSS	99.87 ± 0.10	97.09 ± 0.45	84.08 ± 1.36	78.19 ± 3.74	60.25 ± 2.33	42.50 ± 2.02
	DSS-EXT	99.06 ± 0.32	97.61 ± 0.29	79.54 ± 1.82	76.86 ± 1.93	70.61 ± 1.88	56.10 ± 2.18
	supervised	100.00 ± 0.00	99.78 ± 0.05	99.98 ± 0.00	99.74 ± 0.04	92.02 ± 0.34	95.64 ± 0.09
	1C-Sum	100.00 ± 0.00	97.89 ± 0.54	92.68 ± 3.43	93.69 ± 2.02	83.47 ± 1.09	83.76 ± 2.81

Table 5: Area under the ROC curve for OOD detection with CIFAR 100 as ID. TIN stands for Tiny ImageNet.

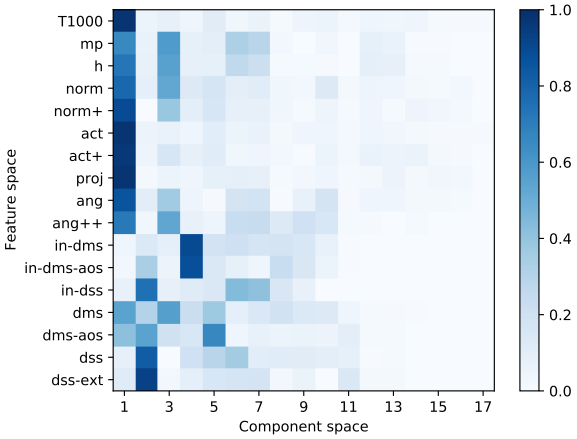
		Gaussian	SVHN	MNIST	fash. MNIST	TIN	LSUN
ResNet 50	ODIN	95.76 ± 5.19	79.33 ± 3.38	81.32 ± 2.85	90.12 ± 0.41	76.40 ± 0.19	72.32 ± 0.91
	T1000	89.86 ± 10.20	84.64 ± 3.11	74.49 ± 2.97	88.95 ± 0.30	77.99 ± 0.10	72.58 ± 0.61
	MP	85.59 ± 9.02	76.65 ± 4.14	69.09 ± 2.66	83.93 ± 0.78	77.34 ± 0.07	73.46 ± 0.17
	H	83.18 ± 9.83	73.58 ± 4.21	64.83 ± 2.69	80.49 ± 0.59	72.86 ± 0.11	72.21 ± 2.45
	NORM	54.94 ± 33.83	71.68 ± 7.59	66.84 ± 4.26	64.45 ± 3.31	51.99 ± 1.12	53.81 ± 1.58
	NORM+	65.88 ± 35.79	78.49 ± 6.22	67.89 ± 2.16	79.23 ± 2.00	65.95 ± 0.60	62.63 ± 1.35
	ACT	89.98 ± 9.92	84.80 ± 3.08	74.66 ± 2.93	88.93 ± 0.30	78.00 ± 0.10	72.51 ± 0.63
	ACT+	90.95 ± 11.23	88.55 ± 2.64	79.42 ± 3.69	88.46 ± 0.29	74.54 ± 0.09	71.80 ± 0.86
	PROJ	89.53 ± 9.93	81.13 ± 4.19	74.02 ± 3.17	88.40 ± 0.28	78.04 ± 0.14	71.18 ± 0.56
	ANG	94.04 ± 2.86	75.70 ± 2.60	70.15 ± 2.99	88.44 ± 0.43	80.53 ± 0.16	72.73 ± 0.12
	ANG++	99.66 ± 0.14	81.42 ± 1.28	82.02 ± 4.09	89.17 ± 0.58	79.26 ± 0.18	75.66 ± 0.62
	IN-DMS	0.01	22.19	88.75	42.44	16.50	32.71
	IN-DMS-AOS	46.01	26.70	98.78	92.11	46.88	48.78
	IN-DSS	0.00	30.31	2.55	11.62	6.87	26.80
	DMS	99.94 ± 0.03	0.25 ± 0.12	13.40 ± 9.11	0.75 ± 0.55	0.46 ± 0.08	21.96 ± 15.56
	DMS-AOS	99.10 ± 0.82	7.03 ± 1.27	77.65 ± 5.84	54.39 ± 4.72	37.60 ± 0.79	40.94 ± 0.32
	DSS	0.00 ± 0.00	4.60 ± 1.34	0.00 ± 0.00	0.00 ± 0.00	0.31 ± 0.22	22.55 ± 15.79
	DSS-EXT	96.67 ± 1.36	95.98 ± 0.36	83.69 ± 2.62	74.09 ± 1.33	52.02 ± 0.68	43.52 ± 0.15
	supervised	100.00 ± 0.00	99.04 ± 0.14	99.98 ± 0.02	99.36 ± 0.24	79.18 ± 0.30	84.88 ± 1.16
	1C-Sum	99.85 ± 0.17	92.97 ± 0.84	84.14 ± 3.56	90.52 ± 0.45	76.93 ± 1.01	70.25 ± 1.43
WideResNet	ODIN	98.48 ± 1.06	81.28 ± 3.49	84.98 ± 2.82	91.10 ± 1.04	77.58 ± 0.42	73.20 ± 2.69
	T1000	94.49 ± 3.78	86.47 ± 2.43	78.51 ± 3.23	89.63 ± 0.92	78.99 ± 0.25	73.05 ± 2.43
	MP	92.03 ± 6.00	77.21 ± 2.06	70.71 ± 2.92	81.93 ± 0.97	76.17 ± 0.17	72.18 ± 1.08
	H	94.37 ± 4.14	80.54 ± 2.23	72.63 ± 3.04	85.29 ± 1.10	78.04 ± 0.11	78.03 ± 3.05
	NORM	67.44 ± 29.49	74.28 ± 5.66	75.02 ± 6.14	79.18 ± 2.71	63.00 ± 2.52	59.63 ± 2.45
	NORM+	80.19 ± 22.36	79.68 ± 4.77	74.43 ± 4.11	85.21 ± 1.60	71.13 ± 1.57	66.82 ± 2.56
	ACT	94.41 ± 3.84	86.74 ± 2.37	78.80 ± 3.17	89.58 ± 0.96	79.00 ± 0.24	72.94 ± 2.42
	ACT+	97.40 ± 2.04	89.09 ± 2.27	82.62 ± 3.83	89.68 ± 0.81	76.47 ± 0.72	70.52 ± 2.77
	PROJ	94.80 ± 3.59	85.51 ± 2.26	79.05 ± 3.61	89.69 ± 0.90	78.90 ± 0.23	73.06 ± 2.13
	ANG	96.22 ± 2.52	82.42 ± 1.30	75.17 ± 2.02	87.68 ± 1.09	79.93 ± 0.33	74.65 ± 1.28
	ANG++	97.02 ± 2.08	84.14 ± 1.29	83.48 ± 3.30	84.41 ± 0.77	75.10 ± 0.21	69.17 ± 1.05
	IN-DMS	0.01	22.19	88.75	42.44	16.50	32.71
	IN-DMS-AOS	46.01	26.70	98.78	92.11	46.88	48.78
	IN-DSS	0.00	30.31	2.55	11.92	6.87	26.80
	DMS	100.00 ± 0.00	84.65 ± 1.38	91.79 ± 1.54	62.49 ± 1.26	39.19 ± 0.54	75.94 ± 30.23
	DMS-AOS	100.00 ± 0.00	6.61 ± 0.81	62.58 ± 5.99	43.24 ± 4.88	40.56 ± 0.81	48.03 ± 0.36
	DSS	99.99 ± 0.00	94.85 ± 0.04	86.68 ± 1.64	82.52 ± 1.22	44.58 ± 0.61	77.31 ± 31.68
	DSS-EXT	85.25 ± 1.53	95.52 ± 0.14	78.13 ± 2.66	75.51 ± 1.68	51.30 ± 0.75	37.70 ± 0.94
	supervised	100.00 ± 0.00	99.22 ± 0.06	99.98 ± 0.02	99.20 ± 0.10	78.40 ± 0.05	81.15 ± 1.64
	1C-Sum	100.00 ± 0.00	95.44 ± 0.99	84.95 ± 4.30	91.06 ± 1.29	77.30 ± 0.32	68.55 ± 2.62
DenseNet 121	ODIN	97.79 ± 2.03	80.72 ± 1.10	75.12 ± 7.12	90.68 ± 1.28	79.22 ± 1.11	74.42 ± 1.55
	T1000	92.50 ± 5.48	87.45 ± 1.47	67.66 ± 6.00	89.77 ± 1.27	80.36 ± 0.78	73.99 ± 1.43
	MP	78.51 ± 13.57	82.17 ± 1.57	67.17 ± 1.57	83.52 ± 0.03	78.33 ± 0.04	74.09 ± 0.88
	H	76.84 ± 13.03	77.78 ± 1.55	61.64 ± 2.34	79.65 ± 0.32	75.32 ± 3.00	74.88 ± 0.96
	NORM	69.60 ± 27.50	58.29 ± 1.80	37.50 ± 14.69	64.73 ± 5.77	60.14 ± 3.49	60.96 ± 2.14
	NORM+	79.09 ± 26.05	69.80 ± 2.88	44.41 ± 14.49	78.80 ± 3.44	70.54 ± 2.26	66.05 ± 2.77
	ACT	92.38 ± 5.56	87.50 ± 1.48	67.77 ± 6.02	89.77 ± 1.25	80.39 ± 0.78	73.97 ± 1.43
	ACT+	93.93 ± 5.11	87.97 ± 0.99	66.48 ± 8.88	88.31 ± 2.40	77.85 ± 1.42	73.97 ± 1.70
	PROJ	92.36 ± 5.32	85.34 ± 2.32	67.46 ± 5.92	89.12 ± 1.12	80.23 ± 0.66	72.53 ± 1.31
	ANG	92.01 ± 7.24	86.73 ± 2.62	77.82 ± 0.67	89.85 ± 0.36	81.26 ± 0.10	72.25 ± 0.35
	ANG++	89.91 ± 13.68	87.72 ± 2.17	84.04 ± 3.14	85.70 ± 0.96	76.23 ± 0.09	71.79 ± 0.25
	IN-DMS	0.01	22.19	88.75	42.44	16.50	42.68
	IN-DMS-AOS	46.01	26.70	98.78	92.11	46.88	48.78
	IN-DSS	0.00	30.31	2.55	11.62	6.87	38.55
	DMS	100.00 ± 0.00	3.41 ± 1.67	6.56 ± 2.87	0.45 ± 0.19	13.92 ± 18.52	35.32 ± 0.33
	DMS-AOS	99.99 ± 0.01	8.99 ± 1.44	51.12 ± 4.77	33.48 ± 2.93	38.62 ± 0.81	50.00 ± 0.67
	DSS	1.84 ± 2.42	6.24 ± 1.04	0.00 ± 0.01	0.00 ± 0.00	14.04 ± 19.82	31.02 ± 1.71
	DSS-EXT	96.60 ± 0.74	95.39 ± 0.67	90.90 ± 0.63	84.26 ± 0.63	50.79 ± 0.32	33.29 ± 1.01
	supervised	100.00 ± 0.00	99.26 ± 0.08	99.98 ± 0.01	99.43 ± 0.09	80.30 ± 0.48	84.84 ± 0.30
	1C-Sum	99.40 ± 0.85	93.23 ± 1.61	79.50 ± 3.08	92.06 ± 1.39	80.10 ± 0.36	72.54 ± 1.78

Table 6: Area under the ROC curve for OOD detection with ImageNet as ID.

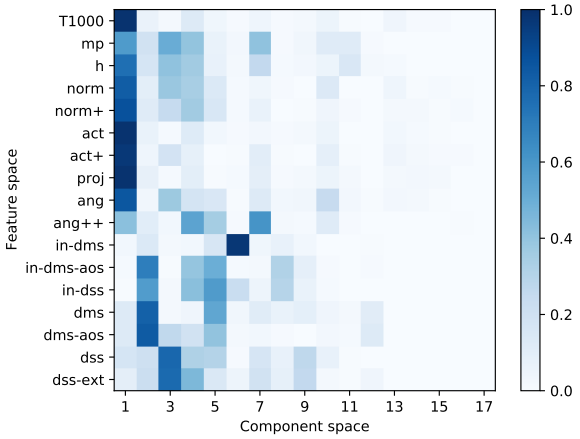
		Gaussian	SVHN	MNIST	fash. MNIST	LSUN	CIFAR 10	CIFAR 100
ResNet 50	ODIN	99.96	99.82	99.73	94.16	80.38	87.66	89.98
	T1000	98.77	98.37	98.03	87.28	78.17	84.23	86.48
	MP	93.87	97.25	91.65	86.99	75.70	81.57	84.75
	H	97.55	98.36	96.16	89.32	77.47	84.36	87.25
	NORM	99.62	95.89	99.53	57.95	62.92	48.98	58.23
	NORM+	99.80	96.35	99.58	63.67	70.37	55.09	63.66
	ACT	98.74	98.37	98.05	87.29	78.15	84.18	86.46
	ACT+	99.94	99.58	99.77	87.29	77.36	83.00	86.89
	PROJ	98.77	98.21	98.43	88.37	77.04	83.59	86.31
	ANG	90.53	93.83	88.94	89.26	74.70	86.40	87.21
	ANG++	99.87	99.56	99.51	98.23	75.03	94.46	96.14
	IN-DMS	26.50	61.58	97.00	68.76	50.00	50.79	56.51
	IN-DMS-AOS	42.46	22.50	98.87	92.60	56.66	38.43	43.72
	IN-DSS	7.52	86.70	55.74	71.10	43.38	54.91	59.55
	DMS	99.92	97.10	99.18	95.69	38.20	86.99	89.04
	DMS-AOS	8.10	16.26	85.79	73.03	52.50	34.10	33.59
	DSS	100.00	98.15	98.19	88.37	31.08	80.06	84.03
	DSS-EXT	100.00	93.17	87.10	72.58	36.36	73.95	77.98
	supervised	100.00	99.98	100.00	99.79	85.18	99.13	99.09
	1C-Sum	100.00	99.43	99.62	93.28	61.76	87.51	91.55
WideResNet	ODIN	100.00	99.91	99.36	96.13	78.42	87.79	89.44
	T1000	99.77	96.59	96.26	88.87	77.00	83.20	85.16
	MP	99.70	95.25	92.17	87.41	77.51	82.89	85.41
	H	99.69	96.96	95.49	89.78	78.90	84.96	87.42
	NORM	80.56	53.96	74.21	40.52	44.73	27.21	31.31
	NORM+	92.88	59.47	80.81	45.66	53.56	32.13	36.98
	ACT	99.77	96.58	96.27	88.88	76.98	83.17	85.14
	ACT+	99.92	97.96	98.25	89.36	70.56	81.72	84.04
	PROJ	99.85	96.67	96.93	90.17	76.64	81.69	84.60
	ANG	98.93	96.16	93.74	92.55	79.65	89.14	90.46
	ANG++	99.98	99.70	99.51	99.15	79.40	96.50	97.12
	IN-DMS	26.50	61.58	97.00	68.76	50.00	50.79	56.51
	IN-DMS-AOS	42.46	22.50	98.87	92.60	56.66	38.43	43.72
	IN-DSS	7.52	86.70	55.74	71.10	43.38	54.91	59.55
	DMS	99.74	95.03	99.92	96.69	48.86	81.34	84.74
	DMS-AOS	4.70	12.07	99.67	83.86	58.68	23.12	26.21
	DSS	99.06	96.37	99.51	96.17	27.21	85.18	87.43
	DSS-EXT	99.99	96.36	92.57	87.23	32.86	84.58	86.41
	supervised	100.00	99.98	100.00	99.98	88.23	99.65	99.54
	1C-Sum	99.94	99.27	99.94	97.56	71.73	88.27	91.35
DenseNet 121	ODIN	100.00	99.54	98.08	92.86	81.90	86.44	88.10
	T1000	99.84	99.02	92.79	88.70	79.93	85.72	87.78
	MP	97.27	97.61	80.44	87.31	76.95	83.16	85.56
	H	99.87	98.79	86.31	90.04	79.01	86.01	88.13
	NORM	99.94	94.53	94.32	49.85	58.91	45.45	57.07
	NORM+	99.97	94.64	95.82	57.67	67.43	51.74	63.24
	ACT	99.85	99.03	92.88	88.71	79.90	85.68	87.76
	ACT+	99.98	99.59	97.33	88.40	77.65	81.84	86.33
	PROJ	99.93	98.94	95.15	88.97	77.85	85.73	87.97
	ANG	95.93	95.93	85.38	90.89	75.49	88.79	88.77
	ANG++	99.96	99.33	97.79	97.94	73.80	93.29	94.81
	IN-DMS	26.50	61.58	97.00	68.76	50.00	50.79	56.51
	IN-DMS-AOS	42.46	22.50	98.87	92.60	56.66	38.43	43.72
	IN-DSS	7.52	86.70	55.74	71.10	43.38	54.91	59.55
	DMS	99.17	90.45	97.55	91.63	52.45	79.22	82.03
	DMS-AOS	0.09	4.23	93.40	76.40	56.95	16.62	18.08
	DSS	100.00	98.92	99.86	97.76	30.66	90.02	92.35
	DSS-EXT	100.00	97.93	94.25	89.08	34.29	88.85	90.68
	supervised	100.00	99.95	99.99	99.87	87.13	98.84	98.69
	1C-Sum	99.99	99.74	99.78	97.63	63.98	92.15	95.00



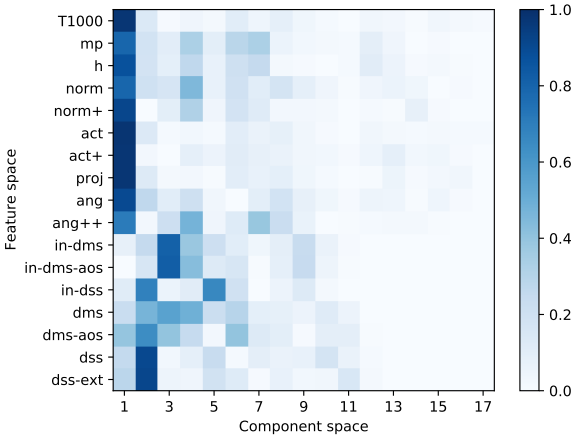
(a) Variance distribution across principal components



(b) Loading analysis on CIFAR 10 (ID set).



(c) Loading analysis on Gaussian noise



(d) Loading analysis on Tiny ImageNet.

Figure 1: PCA analysis of redundancy on CIFAR 10 with ResNet 50. In the loading analyses, pixel on row i and column j expresses the absolute value of the correlation between the ith indicator and the jth component.

C.2. Complementarity/redundancy

We use principal component analysis (PCA) to assess the (linear) redundancy or complementarity of the proposed indicators. For each dataset independently, we created one unsupervised $n \times p$ matrix M_d. $M_d[i, j]$ is value of the jth indicator for the ith sample from dataset d, standardized according to the mean/standard deviation of indicator j on d. We then computed the PCA of each matrices (Figure 1).

Figure 1a shows how the ratio of explained variance evolves with respect to the number of principal components. The first component accounts for 50% of the variance and roughly half of the components are needed to account for the majority (*i.e.* $> 95\%$) of the variance.

Figures 1b-1d show how the components relate to the original indicators. As can be seen, the first component mainly focuses on the optimality-based indicators. The large quantity of variance its explains is somewhat misleading, since more than half of the indicators are so highly correlated themselves. Interestingly, ANG++, H, MP, NORM and NORM+ correlations are spread among two components, mainly, suggesting those might be complementary to the other optimality-based indicators.

To a lesser extend, the second component focuses on the batchnorm indicators. However, several components are needed to fully capture all the batchnorm information. In particular, IN-DMS(-AOS) stand apart from the other indicators. As for IN-DSS, it tends to share its variance with two components. Of the remaining indicators, DSS and DSS-EXT are well correlated and are the main focus of one component. On natural images, DMS and DMS-AOS correlate with several components. On Gaussian noise, though—where they perform well—they stand apart as the second component.

Overall, we observe three main clusters of indicators: (i) the optimality-based ones, (ii) IN- indicators, and (iii) the remaining batchnorm ones. Nevertheless, more than three components are needed to summarize (most of) the variance, since some indicators spread their variance across several components. Although partially redundant, the indicators remain complementary and might help catch different OOD samples.

This analysis suggests that the first few principal components could be used as new indicators that summarize the information contained in the proposed ones. Nevertheless, using the principal components in practice is not trivial, owing to the by-dataset standardization. One could use the same technique as we proposed for 1C-Sum, namely using a (noisy) reference dataset to perform the PCA reduction. We are concerned, however, that this would further bias the detector to perform well mostly on closer OOD dataset to the one used as reference, and degrade the performances in the other cases.

C.3. Model quality

In this section, we would like to investigate how much the quality of the model impacts the quality of OOD detection. To do so, we trained a ResNet 50 on CIFAR 10 and paused the learning at several stages to compute the proposed indicators. We used the accuracy on the training set as criterion to snapshot the performances. Table 7 holds the results with a selection of indicators for several datasets, as well as the test set accuracies.

We can distinguish between three phases. At first, the model is training but not yet overfitting. Around 95% of training accuracy, we see the first evidence of overfitting occurring. Somewhere between 97.5% and 99%, the overfitting is no longer mild. At this point, the gain of accuracy is small and the network decreases its training loss mainly by becoming more confident. A few observations are worth mentionning.

Optimality-based indicators Without surprise, optimality-based indicators suffer from a sub-optimal network. At the first two stages, the proportions of misclassified test samples is relatively high. Since these indicators all rely in some way or another on the predicted class, the poor performances are to be expected. Once reaching 95% of training accuracy, the proportion of misclassified test samples remains stable, and the performances vary less, up to the point where the model becomes overconfident. It does indeed seem easy for the network to push all samples far away from the decision planes in the last latent space, with a regrettable side effect for OOD detection. In this regime, the variance of the results increases as well.

Batchnorm indicators Firstly, note we did not include the IN- indicators because they do not depend on the model. On datasets where they are useful (Gaussian, SVHN), the other batchnorm indicators reach noticeable performances early in the training but keep being refined up till the end. Overconfidence is not a problem for those indicators. They are quite unstable on MNIST and overall worthless against Tiny ImageNet.

Supervised method When supervision is applicable, the model quality plays a much less important role. Except on Tiny ImageNet, the indicator performances with the model trained at 75% of accuracy is already close to its best. Even at the first stage, the linear SVM model is able to discard useless indicators (as against Gaussian, where individual optimality-based indicators perform randomly). As showcases MNIST,

the supervised models also go beyond picking up the best individual indicator. This appraoch is also more robust with respect to overconfidence.

1C-Sum With 1C-Sum, we see that by combining indicators, we can achieve good performances with a less well-trained model (85% of training accuracy, a stage more on Tiny ImageNet). At that stage, this indicator is already performing better than most other (sample-free) indicator at their peak. Unlike the supervised method, this approach suffers somewhat from overconfidence.

CIFAR 10 vs. CIFAR 100 At 75% of training accuracy, the network has comparable test accuracy as optimal-ones on CIFAR 100 (Table 3). Against the Gaussian dataset, optimality-based indicators perform much better in the case of CIFAR 100. Against SVHN and Tiny ImageNet, the performance are similar in both cases. Against MNIST, indicators work better in the case of CIFAR 10.

Discussion As expected, optimality-based indicators suffer from a sub-optimal model. They also suffer from an over-confident network. 1C-Sum follows the same trend but depends less on the model optimality. It is also less impacted by overconfidence. Batchnorm indicators are less predictable but seem to be also impacted by the model quality when they are useful. Some supervision can compensate for the lack of training.

C.4. Misclassification detection

In this section, we investigate whether wrongly rejected ID samples correspond to misclassified ones. We performed two experiments.

Error detection The first experiment consist in using the proposed indicators to detect misclassifications: we only look at the training set of the base task and label as positive the samples for which the network makes an classification error; samples for which the model is correct are labeled as negative. Table 8 shows the area under the ROC curve for detecting these positive samples. As can be seen, not all indicators are equal in this respect. Indicator appropriateness is stable across architectures and datasets.

Optimality-based indicators are clearly best suited: a well-optimized network should lower its confidence when making a mistake. Among those, MP and H stand out, then comes ANG, followed by ACT, PROJ and T1000. For this task, the positive variant are less adequate. T1000 always outperforms ODIN; the adversarial perturbation will lower the network confidence blurring the separation between positive and negative samples.

Batchnorm indicators are not suited for the task, suggesting that the network mistakes are not due to statistical outliers. As for the aggregating indicator, 1C-Sum, it performs adequately, although simpler indicators work better.

Even though the networks make more mistakes on CIFAR 100, detecting them is a harder challenge. Whether this is caused by a less well-performing model, or by other factors (such the number of classes which might spread the predictions more across classes) is not clear.

Joint OOD and misclassification detection In the second experiment, we are considering both OOD samples and misclassification as the positive class. In other words, are considered negative samples only those of the base task for which the network predicts the correct class. Although misclassifications cannot count as OOD samples *per se*, it might be more interesting in practice to reject those as well.

Table 9 shows the average (over the OOD datasets—the same as for the other experiments) improvement in auroc when tackling the joint task rather than OOD detection only. It is confirmed that wrongly rejected ID samples are partly due to misclassified one, in the case of optimality-based indicators. Indeed, they benefit from a raise of auroc, which is more pronounced with CIFAR 100, where there are much more classification errors.

We also see that this is not true of batchnorm indicators, albeit IN-DMS and IN-DSS see a small improvement. This suggests that misclassified samples are not necessarily statistically off compared to other ID samples.

Interestingly, 1C-Sum benefits slightly from the joint task, even though it incorporates batchnorm indicators.

The previous analyses highlighted that indicators good at detecting misclassifications might differ from those best at OOD detection. However, when the network performs well, misclassified samples should be negligible. This is confirmed by table 10, which displays the average (across OOD datasets—the same as for the other experiments) top rank for each indicator at this joint task. On CIFAR 10, the relative order of indicators is mostly unchanged. On CIFAR 100, MP and H are better positionned in the ranking, although the number of mistakes might be too low for them to outperforms the best indicators.

Overall, it does seem that some of the wrongly rejected ID samples are also misclassified. 1C-Sum remains the best bet to tackle OOD detection, possibly jointly with misclassification rejection, at least in the absence of data.

Table 7: Model quality and its impact on OOD detection. A ResNet 50 was trained on CIFAR 10 and paused when reaching some training set accuracy threshold (first row) to examine how the features perform. The metric is the area under the ROC curve. Coloring reflects the 50% of *overall* best results per dataset.

	Train accuracy (%)	75	85	95	97.5	99	full
	Test accuracy (%)	77.15 ± 1.49	84.99 ± 0.36	92.41 ± 0.49	93.49 ± 0.07	93.58 ± 0.06	94.11 ± 0.25
Gaussian	T1000	49.23 ± 13.93	89.71 ± 2.95	96.95 ± 3.31	97.28 ± 1.61	95.41 ± 4.59	83.17 ± 9.00
	MP	41.31 ± 15.76	84.84 ± 5.74	95.06 ± 5.27	95.23 ± 2.73	93.31 ± 5.37	89.27 ± 4.90
	H	42.82 ± 16.58	87.59 ± 4.08	95.65 ± 4.80	96.32 ± 2.36	94.09 ± 5.33	89.05 ± 5.03
	NORM	47.98 ± 13.03	83.74 ± 1.46	96.57 ± 3.76	96.55 ± 1.80	94.61 ± 5.47	53.96 ± 33.02
	ACT	49.50 ± 13.95	89.51 ± 3.40	96.87 ± 3.43	97.28 ± 1.72	95.48 ± 4.50	83.34 ± 9.02
	PROJ	49.47 ± 14.99	89.69 ± 4.11	96.57 ± 3.97	97.17 ± 2.05	95.63 ± 4.02	85.53 ± 8.09
	ANG	51.94 ± 14.55	89.41 ± 5.93	95.85 ± 4.38	96.20 ± 2.46	94.73 ± 3.90	91.78 ± 2.79
	DMS-AOS	84.91 ± 11.58	85.95 ± 4.30	86.66 ± 4.07	92.49 ± 3.40	95.01 ± 2.37	99.25 ± 0.48
	DSS-EXT	94.71 ± 5.24	96.60 ± 0.61	98.64 ± 0.55	98.77 ± 0.26	98.17 ± 0.79	98.24 ± 0.61
	supervised	99.99 ± 0.01	100.00 ± 0.00	100.00 ± 0.01	100.00 ± 0.00	99.99 ± 0.01	100.00 ± 0.00
	1C-Sum	81.41 ± 9.90	99.54 ± 0.13	99.88 ± 0.16	99.95 ± 0.06	99.81 ± 0.21	97.84 ± 2.70
SVHN	T1000	82.65 ± 3.65	94.68 ± 1.68	97.17 ± 1.05	97.32 ± 0.39	96.49 ± 0.32	93.14 ± 3.05
	MP	80.17 ± 4.68	92.60 ± 1.17	94.89 ± 1.59	94.73 ± 0.72	94.32 ± 1.27	91.89 ± 1.30
	H	81.69 ± 3.45	94.36 ± 1.35	96.36 ± 1.33	96.00 ± 0.66	95.19 ± 1.14	92.51 ± 1.46
	NORM	85.57 ± 3.88	95.52 ± 2.72	97.27 ± 1.90	97.58 ± 0.65	94.55 ± 1.83	85.46 ± 10.89
	ACT	83.30 ± 3.53	94.94 ± 1.61	97.24 ± 1.05	97.42 ± 0.39	96.58 ± 0.30	93.32 ± 2.95
	PROJ	81.85 ± 4.51	94.84 ± 1.35	97.34 ± 0.96	97.40 ± 0.38	96.68 ± 0.29	94.01 ± 2.42
	ANG	75.08 ± 5.28	90.84 ± 0.33	95.32 ± 0.29	94.63 ± 0.45	94.36 ± 1.19	93.41 ± 0.09
	DMS-AOS	18.21 ± 6.46	11.38 ± 10.15	2.72 ± 1.27	2.46 ± 0.68	3.92 ± 1.70	4.72 ± 2.26
	DSS-EXT	95.32 ± 0.83	97.80 ± 0.45	98.40 ± 0.22	98.20 ± 0.33	97.88 ± 0.18	97.70 ± 0.34
	supervised	98.94 ± 0.23	99.50 ± 0.23	99.72 ± 0.01	99.67 ± 0.04	99.65 ± 0.04	99.75 ± 0.05
	1C-Sum	89.12 ± 5.17	98.49 ± 0.59	99.00 ± 0.22	98.98 ± 0.20	98.71 ± 0.17	97.83 ± 0.95
MNIST	T1000	83.64 ± 4.91	89.35 ± 1.44	94.70 ± 0.18	94.88 ± 0.34	94.88 ± 2.18	94.81 ± 0.78
	MP	75.71 ± 6.21	82.44 ± 1.47	90.03 ± 0.77	89.85 ± 1.00	89.47 ± 2.48	90.76 ± 0.65
	H	78.97 ± 6.44	85.59 ± 1.33	91.76 ± 0.64	91.09 ± 0.94	90.41 ± 2.67	91.40 ± 0.62
	NORM	76.20 ± 1.94	88.82 ± 5.56	94.92 ± 1.60	95.17 ± 1.58	97.36 ± 0.93	92.28 ± 4.92
	ACT	83.71 ± 4.25	89.44 ± 1.53	94.73 ± 0.19	94.92 ± 0.34	94.93 ± 2.19	94.90 ± 0.70
	PROJ	83.74 ± 4.24	90.42 ± 1.73	95.30 ± 0.22	95.54 ± 0.30	95.59 ± 1.75	95.61 ± 0.40
	ANG	83.70 ± 5.68	88.18 ± 0.57	93.65 ± 1.14	93.77 ± 1.50	91.93 ± 2.90	94.15 ± 0.60
	DMS-AOS	80.49 ± 10.05	64.11 ± 10.43	57.98 ± 2.28	71.07 ± 3.05	65.48 ± 5.40	81.12 ± 9.04
	DSS-EXT	49.66 ± 2.26	68.02 ± 3.52	73.17 ± 1.54	70.39 ± 5.40	68.64 ± 3.40	66.93 ± 1.88
	supervised	99.99 ± 0.01	99.99 ± 0.01	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
	1C-Sum	85.02 ± 7.01	93.76 ± 1.89	96.37 ± 0.40	96.91 ± 0.51	96.53 ± 1.73	96.47 ± 1.58
Tiny ImageNet	T1000	77.54 ± 1.42	83.33 ± 1.67	89.57 ± 0.05	90.17 ± 0.18	89.46 ± 0.36	88.70 ± 1.23
	MP	73.90 ± 1.42	80.37 ± 1.76	87.05 ± 0.17	87.55 ± 0.18	87.20 ± 0.26	87.05 ± 0.61
	H	75.72 ± 1.47	88.59 ± 1.49	88.23 ± 0.12	91.02 ± 0.11	90.06 ± 0.27	87.52 ± 0.67
	NORM	74.83 ± 1.64	80.70 ± 2.21	87.55 ± 0.17	87.16 ± 0.34	83.57 ± 1.55	80.19 ± 4.27
	ACT	77.27 ± 1.46	83.09 ± 1.79	89.47 ± 0.09	90.14 ± 0.19	89.47 ± 0.38	88.77 ± 1.18
	PROJ	77.32 ± 1.36	82.91 ± 1.80	89.34 ± 0.10	89.93 ± 0.10	89.30 ± 0.38	88.61 ± 1.26
	ANG	75.32 ± 1.35	81.02 ± 1.16	87.99 ± 0.15	88.81 ± 0.12	88.47 ± 0.23	88.35 ± 0.51
	DMS-AOS	34.29 ± 2.51	29.05 ± 2.07	22.83 ± 1.20	22.83 ± 1.01	25.31 ± 1.19	25.25 ± 2.65
	DSS-EXT	64.33 ± 0.55	65.97 ± 1.83	68.01 ± 0.55	68.29 ± 0.99	67.92 ± 0.39	66.84 ± 0.88
	supervised	80.44 ± 1.44	85.22 ± 1.50	90.71 ± 0.27	90.89 ± 0.14	90.56 ± 0.17	90.82 ± 0.45
	1C-Sum	76.12 ± 2.90	82.84 ± 2.39	89.33 ± 0.27	89.95 ± 0.26	88.92 ± 0.78	88.86 ± 0.79

C.5. Semantic anomalies

Recently, [1] proposed to distinguish OOD detection between statistical and semantic anomalies detection. A statistical shift occurs for instance when the network is presented with a sample for which it knows the class but under new lighting conditions. In opposition, a semantic anomaly is a sample of an unknown class (at training

Table 8: Error detection. Indicator performance (area under the ROC curve) for misclassification prediction.

	CIFAR 10			CIFAR 100		
	ResNet 50	WideResNet	DenseNet 121	ResNet 50	WideResNet	DenseNet 121
ODIN	85.63 ± 2.92	83.38 ± 3.84	75.17 ± 1.77	76.86 ± 0.14	78.66 ± 1.13	78.55 ± 1.08
T1000	88.99 ± 1.39	88.60 ± 0.35	85.70 ± 1.85	79.13 ± 0.32	79.78 ± 0.67	79.96 ± 0.27
MP	93.13 ± 0.59	92.96 ± 0.39	92.58 ± 0.28	86.54 ± 0.47	86.47 ± 0.19	87.32 ± 0.35
H	93.12 ± 0.59	92.88 ± 0.38	92.47 ± 0.27	86.19 ± 0.52	86.29 ± 0.17	86.87 ± 0.27
NORM	74.79 ± 5.41	70.38 ± 7.18	50.37 ± 3.27	52.18 ± 0.72	62.44 ± 2.72	59.72 ± 3.18
NORM+	81.55 ± 3.41	80.04 ± 3.75	66.54 ± 3.13	65.77 ± 0.37	70.33 ± 1.86	69.47 ± 1.64
ACT	89.11 ± 1.33	88.63 ± 0.40	85.86 ± 1.80	79.18 ± 0.32	79.80 ± 0.65	80.02 ± 0.28
ACT+	87.35 ± 1.99	84.90 ± 2.30	78.52 ± 2.92	74.06 ± 0.43	76.83 ± 1.17	76.94 ± 1.15
PROJ	88.97 ± 1.37	88.74 ± 0.12	86.09 ± 1.80	80.56 ± 0.28	79.74 ± 0.56	80.80 ± 0.13
ANG	90.73 ± 0.52	91.79 ± 0.25	92.03 ± 0.58	84.06 ± 0.35	81.45 ± 0.20	82.14 ± 1.21
ANG++	90.48 ± 0.77	86.99 ± 0.73	86.96 ± 2.15	82.24 ± 0.29	76.84 ± 0.62	77.35 ± 0.60
DMS	29.77 ± 4.79	37.70 ± 1.88	34.21 ± 1.89	38.76 ± 0.53	39.46 ± 0.64	39.69 ± 0.97
DMS-AOS	31.05 ± 1.91	38.47 ± 1.51	33.11 ± 0.97	39.47 ± 0.61	41.90 ± 0.29	41.12 ± 0.75
DSS	40.39 ± 0.34	41.14 ± 2.19	45.40 ± 1.12	43.63 ± 1.06	44.46 ± 0.34	41.59 ± 1.29
DSS-EXT	53.46 ± 0.63	49.95 ± 1.84	56.76 ± 0.91	49.13 ± 0.27	49.57 ± 0.38	47.81 ± 0.39
1C-SUM	89.42 ± 2.64	86.24 ± 2.97	85.02 ± 1.87	78.95 ± 1.03	79.51 ± 0.89	81.82 ± 0.62

Table 9: Average auroc score improvement when tackling the joint task of OOD and misclassification detection.

	CIFAR 10			CIFAR 100		
	ResNet 50	WideResNet	DenseNet 121	ResNet 50	WideResNet	DenseNet 121
ODIN	1.30 ± 0.68	0.87 ± 0.51	0.60 ± 0.79	4.39 ± 2.00	4.63 ± 2.53	4.49 ± 2.22
T1000	1.71 ± 0.93	1.22 ± 0.60	1.29 ± 0.51	5.06 ± 2.74	4.99 ± 2.97	4.88 ± 2.59
MP	2.64 ± 0.60	2.22 ± 0.59	2.15 ± 0.43	8.01 ± 2.35	8.10 ± 3.16	8.42 ± 2.91
H	2.47 ± 0.65	1.96 ± 0.66	1.99 ± 0.47	9.01 ± 2.29	6.73 ± 3.05	9.25 ± 2.62
NORM	1.05 ± 1.20	0.21 ± 0.80	0.02 ± 0.79	-0.07 ± 2.62	2.19 ± 3.60	1.62 ± 2.94
NORM+	1.40 ± 1.16	0.65 ± 0.70	0.75 ± 0.71	2.78 ± 2.84	3.47 ± 3.26	3.40 ± 3.23
ACT	1.71 ± 0.94	1.20 ± 0.61	1.29 ± 0.52	5.06 ± 2.71	4.99 ± 2.95	4.90 ± 2.58
ACT+	1.33 ± 0.87	0.82 ± 0.69	0.99 ± 0.67	3.21 ± 2.74	3.63 ± 3.37	3.85 ± 2.70
PROJ	1.64 ± 0.89	1.21 ± 0.62	1.29 ± 0.55	5.67 ± 2.76	4.99 ± 2.99	5.31 ± 2.62
ANG	1.81 ± 0.47	1.83 ± 0.51	1.40 ± 0.56	6.48 ± 2.87	5.75 ± 3.29	5.61 ± 2.86
ANG++	1.25 ± 0.92	1.63 ± 0.80	1.19 ± 0.92	4.61 ± 3.53	4.26 ± 3.36	4.10 ± 3.08
DMS	-1.48 ± 1.16	-1.12 ± 1.05	-1.18 ± 1.03	-1.27 ± 4.83	-5.49 ± 4.35	-1.77 ± 4.21
DMS-AOS	-1.38 ± 1.22	-1.10 ± 1.09	-1.40 ± 1.00	-3.82 ± 3.53	-3.74 ± 3.87	-3.39 ± 3.44
DSS	-1.08 ± 1.02	-0.96 ± 1.02	-0.71 ± 0.96	2.56 ± 2.98	-4.27 ± 4.50	1.69 ± 2.48
DSS-EXT	-0.41 ± 0.91	-0.52 ± 0.83	-0.19 ± 0.88	-1.81 ± 3.18	-1.20 ± 2.80	-1.77 ± 3.37
1C-SUM	1.20 ± 0.93	0.73 ± 0.81	0.85 ± 0.87	3.25 ± 3.56	3.37 ± 4.09	4.01 ± 3.45

time) but from a close statistical distribution. The authors illustrate this by using a left-out class of CIFAR 10 to define OOD samples with respect to a model trained from all other classes. Note that most of the time, OOD data are both statistically and semantically different. For instance, MNIST is neither semantically nor statistically close to CIFAR 10.

Table 11 contains some results relating to the problem of semantic anomalies. We split the classes of CIFAR 10 in two halves. The first half, h1, contains the airplane, bird, car, cat and deer classes (two vehicles, three animals). The second half, h2, contains the dog, horse, monkey, ship and truck classes (two vehicles and three animals). We learned a ResNet 50 model on each half separately and use the remaining one as OOD data (accuracies are 95.4 ± 0.6 and 97.6 ± 0.1 for the first and second halves respectively).

Interestingly, one OOD task is harder than the other (whereas the accuracy on the ID tasks are roughly equivalent). Without surprise the batchnorm features are inadequate for semantic anomalies. The baseline indicators are hard to outperform in this setting. Even ANG++, which sometimes excel, is hard pressed here. The proposed 1C-SUM indicator does not perform so well either. This is mainly due to the batchnorm features dragging it down. When removed, 1C-SUM* performance increases significantly.

The semantic anomaly detection subproblem is very

Table 10: Average top ranking for the joint task of misclassification and OOD detection

	CIFAR 10			CIFAR 100		
	ResNet 50	WideResNet	DenseNet 121	ResNet 50	WideResNet	DenseNet 121
ODIN	6.83	7.00	9.67	5.83	5.50	5.83
T1000	6.83	5.33	8.33	6.33	6.33	6.00
MP	10.17	8.83	5.67	6.67	8.17	7.00
H	9.00	7.33	4.67	9.17	6.17	9.33
NORM	13.67	12.83	16.83	13.67	14.83	13.67
NORM+	12.17	10.17	14.17	12.50	12.83	12.50
ACT	5.83	4.00	7.33	6.00	6.33	5.17
ACT+	6.67	6.83	11.67	7.83	7.00	8.33
PROJ	6.00	4.00	6.33	6.67	5.83	6.67
ANG	6.17	6.83	2.17	5.33	5.67	3.33
ANG++	7.00	11.83	5.83	3.00	8.83	8.00
DMS	14.33	13.83	12.33	17.00	13.83	16.83
DMS-AOS	16.17	17.67	17.50	14.67	17.33	14.33
DSS	11.83	13.00	11.67	18.67	11.33	18.50
DSS-EXT	11.33	13.17	10.83	10.33	13.33	9.50
1C-SUM	2.67	3.83	3.83	3.83	3.67	3.00

challenging and might benefit most from having access to ID data.

Table 11: Semantic anomaly detection. A ResNet 50 was trained on half the classes of CIFAR 10 while the remaining classes are used as OOD set (h1 contains the five first classes, h2 the second half). The metric used is the Area under the ROC curve. Coloring reflects the 50% best results (darker is better). Note that 1C-SUM* only incorporates the non-batchnorm features of 1C-SUM.

Indicator	h1 (id) / h2 (ood)	h2 (id) / h1 (ood)
T1000	77.15 ± 1.07	88.02 ± 0.63
MP	77.77 ± 0.81	87.59 ± 0.61
H	77.88 ± 0.81	87.76 ± 0.62
NORM	69.26 ± 0.86	83.49 ± 0.76
NORM+	73.24 ± 1.37	86.14 ± 0.79
ACT	77.04 ± 1.12	87.91 ± 0.65
ACT+	75.17 ± 1.02	86.96 ± 0.79
PROJ	75.88 ± 1.08	87.87 ± 0.68
ANG	76.84 ± 0.44	87.31 ± 0.63
ANG++	74.48 ± 0.82	83.40 ± 0.11
IN-NOTA	45.12	55.28
IN-DMS	43.87	56.43
IN-DMS-AOS	53.34	46.95
IN-DSS	41.23	59.21
NOTA	50.00 ± 0.00	49.54 ± 0.66
DMS	32.34 ± 1.07	36.97 ± 3.71
DMS-AOS	44.17 ± 0.95	37.22 ± 0.64
DSS	43.13 ± 1.78	57.45 ± 1.94
DSS-EXT	48.73 ± 1.63	58.92 ± 0.41
1C-SUM	71.30 ± 0.99	84.25 ± 0.73
1C-SUM*	76.17 ± 1.15	87.70 ± 0.60

D. Selected indicator distributions

In this section, we would like to share some distributions of the indicators. Figure 2 displays the distribution for some of bounded indicators. In theory, those are the easiest to threshold without data. In practice, setting the cut points without data is challenging. Interestingly, in an application where rejecting ID samples is less of a problem, MP and H seem good candidate to minimize the OOD acceptance rate. This also explains why they benefit so much from rejecting misclassified samples: ID samples is their main source of mistakes—reducing the number of such samples improves the auroc score.

Figure 3 displays some distributions for unbounded indicators. It does seem that pinpointing where to place the threshold is quite hard.

Figures 4 and 5 focuses on batchnorm indicators. As one can see, these indicators are of limited use in most cases.

Finally, Figure 6 displays the distribution of 1C-Sum in various settings. Without any prior knowledge, placing the optimal threshold is, once more, challenging. The dependency on the network might be more important than the one on the ID task (ResNet 50 for ImageNet is slightly different than for CIFAR 10/100). We leave the evaluation of transferability/meta-learning of the threshold as future work.

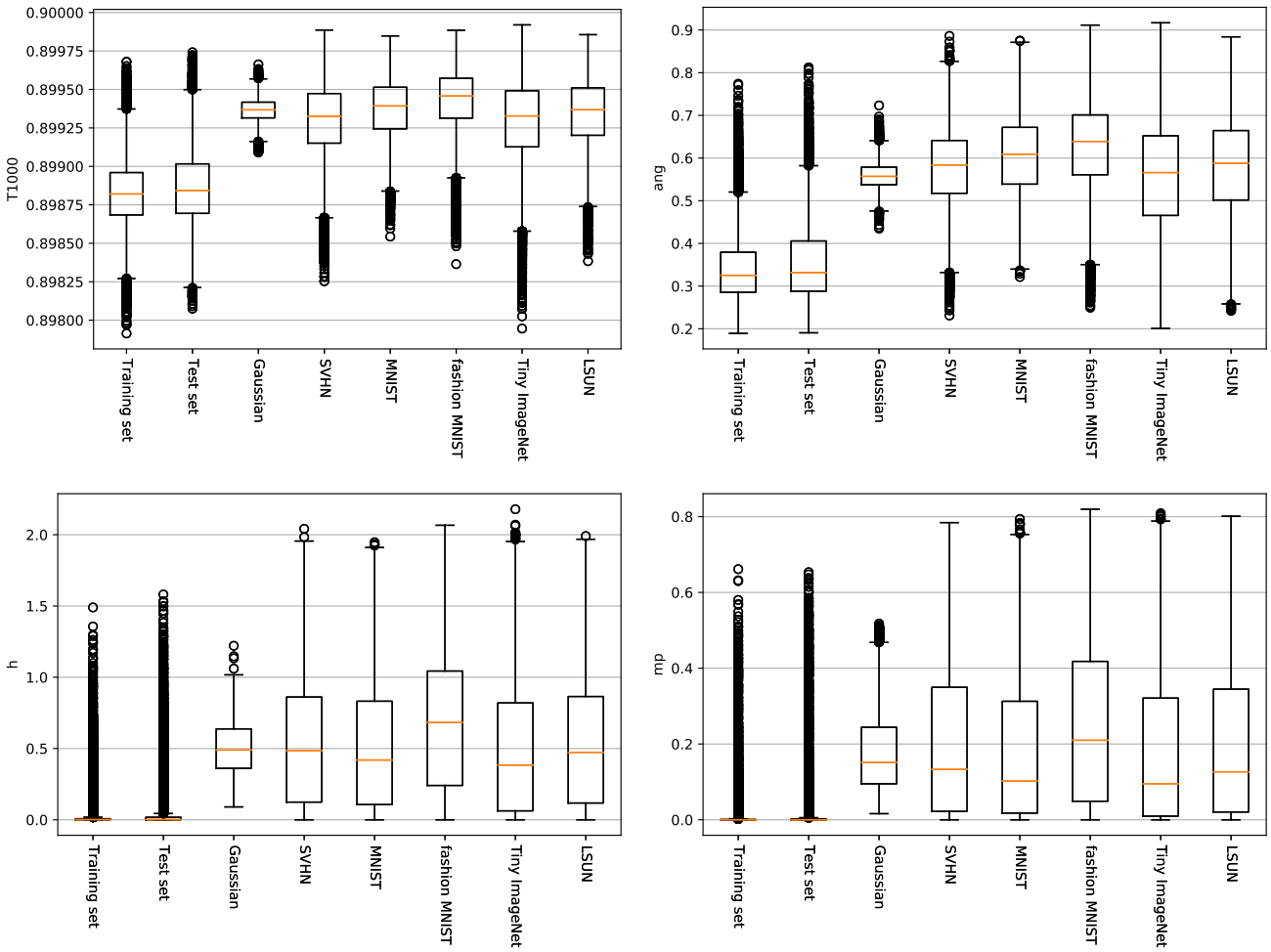


Figure 2: Bounded indicator distributions established with CIFAR 10 as ID task on ResNet 50.

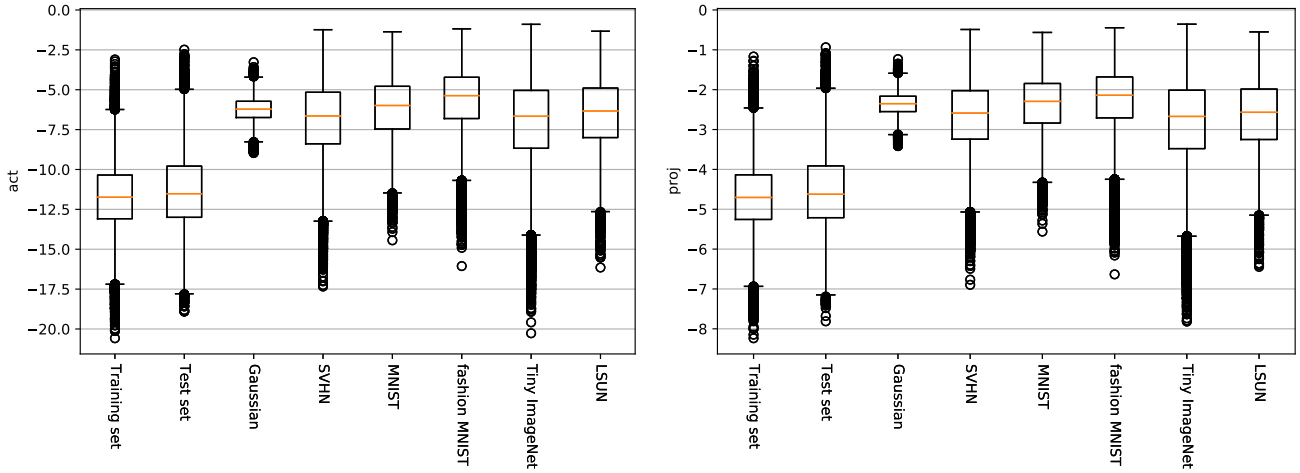


Figure 3: Unbounded indicator distributions established with CIFAR 10 as ID task on ResNet 50.

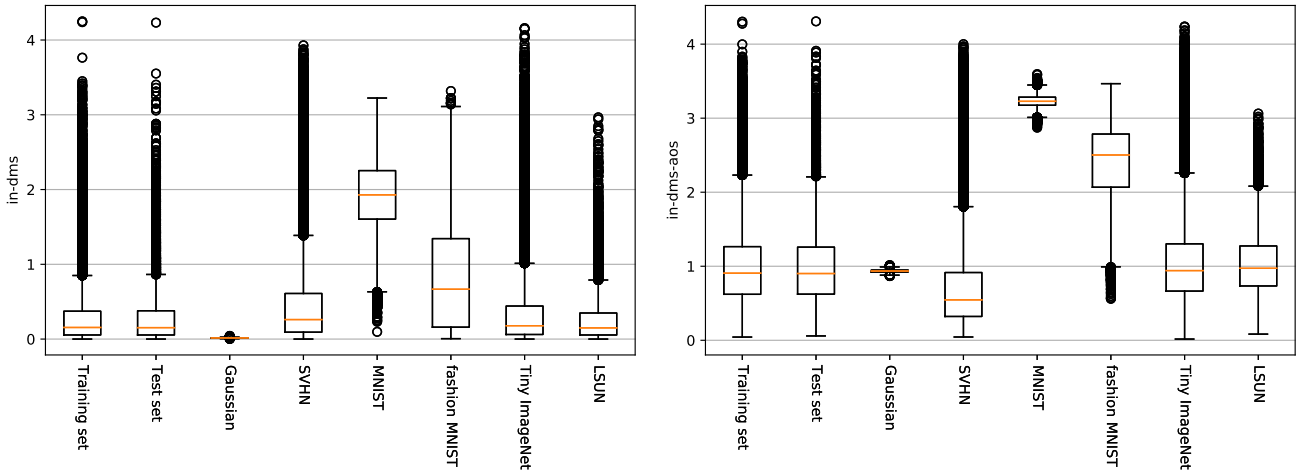


Figure 4: IN- indicator distributions established with CIFAR 10 as ID task on ResNet 50.

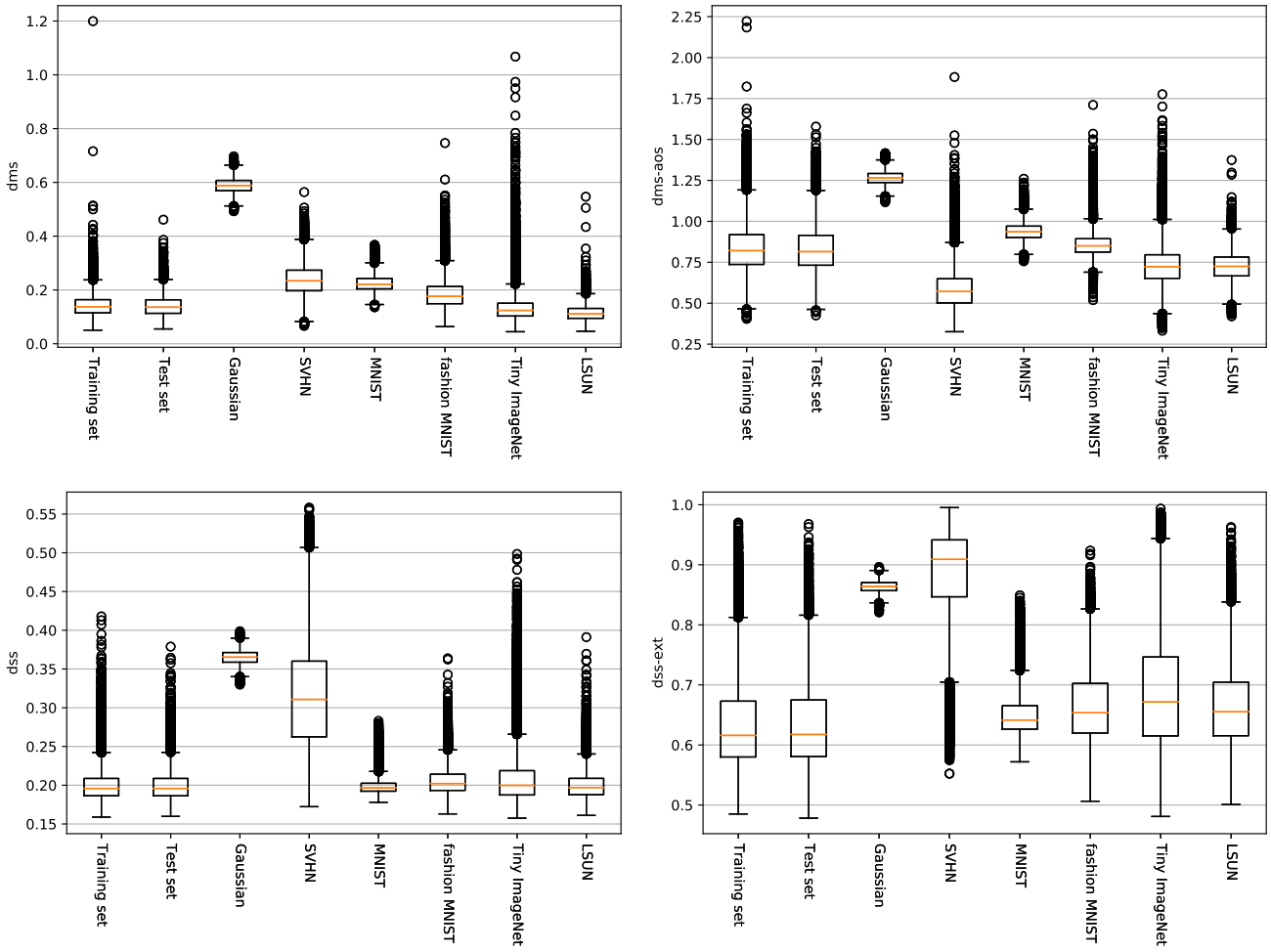
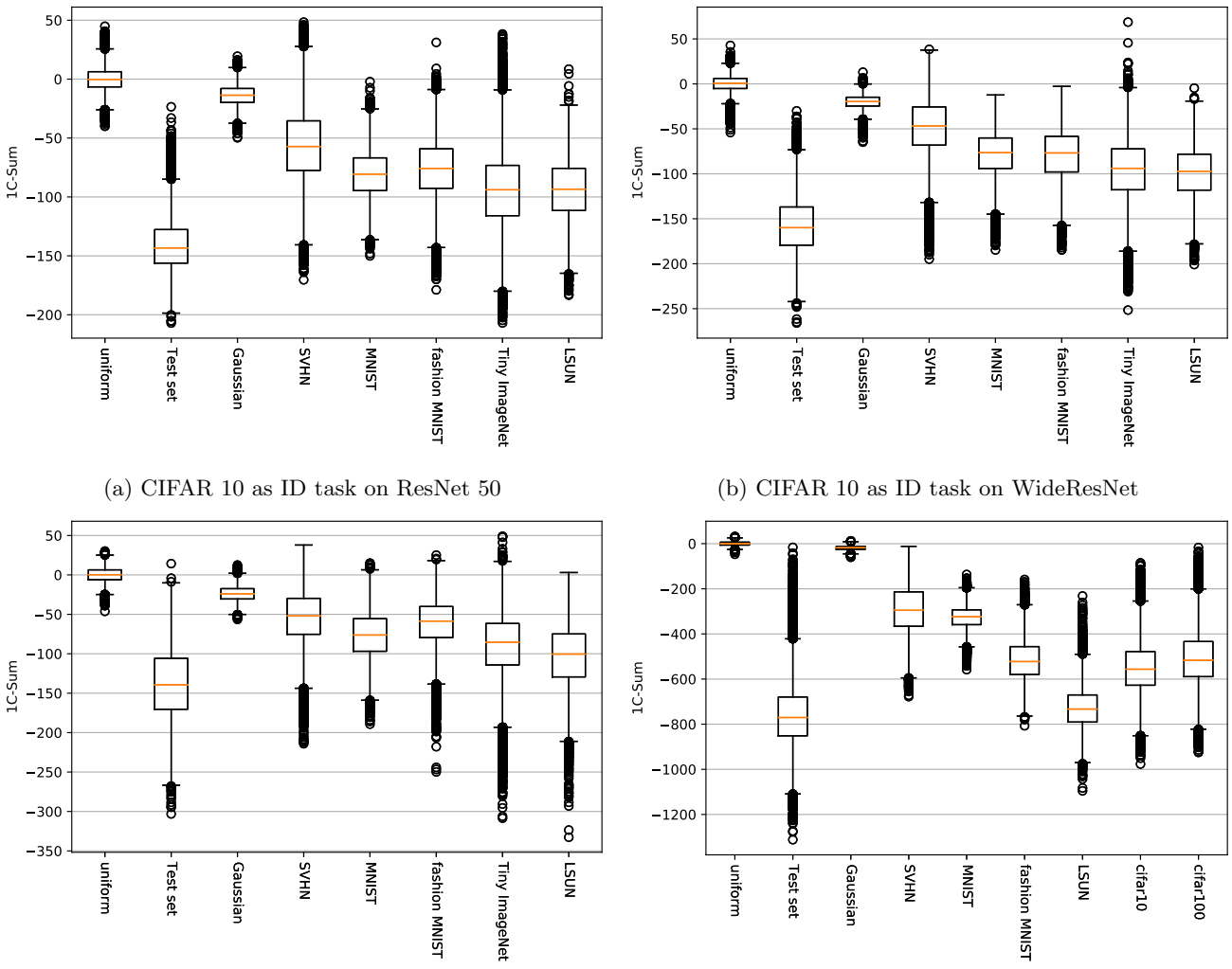


Figure 5: Batchnorm indicator distributions established with CIFAR 10 as ID task on ResNet 50.



(a) CIFAR 10 as ID task on ResNet 50

(b) CIFAR 10 as ID task on WideResNet

(c) CIFAR 100 as ID task on ResNet 50

(d) ImageNet as ID task on ResNet 50

Figure 6: 1C-Sum indicator distributions.

References

[1] Faruk Ahmed and Aaron C. Courville. Detecting semantic anomalies. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 3154–3162. AAAI Press, 2020. 12

[2] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995. 4

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[5] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 3

[6] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 3

[7] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 3

[8] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In Richard C. Wilson, Edwin R. Hancock, and William A. P. Smith, editors, *Proceedings of the British Machine Vision Conference 2016, BMVC 2016, York, UK, September 19-22, 2016*. BMVA Press, 2016. 3