Revisiting Out-of-Distribution Detection: A Simple Baseline is Surprisingly Effective

Julian Bitterwolf [1] **Alexander Meinke** [1] **Maximilian Augustin** [1] **Matthias Hein** [1]

Abstract

It is an important problem in trustworthy machine learning to recognize out-of-distribution (OOD) inputs which are inputs unrelated to the in-distribution task. The goal of this paper is to identify common objectives as well as the identification of the implicit scoring functions of different OOD detection methods. In particular, we show that binary discrimination between in- and (different) out-distributions in combination with classifier outputs is equivalent to several differing formulations of the OOD detection problem and that this binary discriminator reaches an OOD detection performance similar to that of Outlier Exposure.

1. Introduction

While deep learning has significantly improved performance in many application domains, there are serious concerns for using deep neural networks in applications which are of safety-critical nature. With one major problem being overconfident predictions [15, 6, 5] for images not belonging to the classes of the actual task. Here, one distinguishes between far out-distribution data, e.g. different forms of noise or completely unrelated tasks like CIFAR-10 vs. SVHN, and close out-distribution data which can for example occur in related image classification tasks where the semantic structure is very similar e.g. CIFAR-10 vs. CIFAR-100. Both are important, but close out-distribution data is the more difficult problem with potentially fatal consequences: in an automated diagnosis system we want that the system recognizes that an unseen disease is present rather than assigning high confidence into a known class leading to fatal treatment decisions. One of the most effective methods for out-distribution detection is Outlier Exposure [7] and work building upon it [4, 12, 13, 1, 16, 17] where a classifier is trained on the in-distribution task and one enforces low confidence during training on a large and diverse set of out-distribution images. Recently, NTOM [4] has achieved excellent results for detecting far out-distribution data by adding a background class to the classifier which is trained on samples that show a desired hardness for the model. Even though it has been claimed that new approaches outperform [7], up to our knowledge this has not been shown consistently across different and challenging test out-distribution datasets (including close and far out-distribution datasets).

Our main contributions are that we show that several OOD detection approaches are equivalent to a binary discriminator between in- and out-distribution when analyzing the rankings induced by the Bayes optimal classifier/density. Additionally, we derive the implicit scoring functions for Outlier Exposure [7] and using an additional background class for the out-distribution. Finally, we show that when training the binary discriminator between in- and out-distribution together with a standard classifier on the in-distribution in a shared fashion, the binary discriminator reaches state-of-the-art OOD detection performance. However, while we identify that a simple baseline is competitive with the state-of-the-art, the main aim of this paper is a better understanding of the key components of different OOD detection methods and to identify the key properties which lead to SOTA OOD detection performance. All of our findings are supported by extensive experiments on CIFAR-10 and CIFAR-100 with evaluation on various challenging out-distribution test datasets.

2. Bayes-optimal Behaviour for OOD Detection

The OOD problem In order to make rigorous statements about the OOD detection problem we first have to provide the mathematical basis for doing so. We assume that we are given an in-distribution $p(x|i)$ and potentially also a *training* out-distribution $p(x|o)$. At this particular point no labeled data is involved, so both of them are just distributions over X. For simplicity we assume in the following that they both have a density with respect to the Lebesgue measure on $X = [0, 1]^d$. We assume that in practice we get samples

[1]Department of Science, University of Tübingen. Correspondence to: Julian Bitterwolf <julian.bitterwolf@uni-tuebingen.de>.

Presented at the ICML 2021 Workshop on Uncertainty and Robustness in Deep Learning.,

from the mixture distribution

$$\begin{aligned} p(x) &= p(x|i)p(i) + p(x|o)p(o) \\ &= p(x|i)p(i) + p(x|o)(1-p(i)), \end{aligned} \tag{1}$$

where $p(i)$ is the probability that we expect to see in-distribution samples. In order to make the optimal decision between in-and out-distribution for a given x we consider

$$p(i|x) = \frac{p(x|i)p(i)}{p(x)} = \frac{p(x|i)p(i)}{p(x|i)p(i) + p(x|o)p(o)}, \tag{2}$$

which is defined for all $x \in [0,1]^d$ with $p(x) > 0$. If the training out-distribution is the test out-distribution, this is optimal. We would like the approach to generalize to other test out-distributions, which depends on the particular training out-distribution $p(x|o)$. Note that as $p(i|x)$ is only well-defined for all x with $p(x) > 0$, it is reasonable to choose a distribution with support everywhere in $[0,1]^d$, that is $p(x|o) > 0$ for all $x \in [0,1]^d$.

Optimal prediction of a binary discriminator between in- and out-distribution We consider a binary discriminator with model parameters θ between in- and (training) out-distribution, where $\hat{p}_\theta(i|x)$ is the predicted probability for the in-distribution. Under the assumption that $p(i)$ is the probability for in-distribution samples and using logistic loss the expected loss becomes:

$$\begin{aligned} \min_\theta \quad & p(i) \mathop{\mathbb{E}}_{x\sim p(x|i)} [-\log \hat{p}_\theta(i|x)] \\ & + \; p(o) \mathop{\mathbb{E}}_{x\sim p(x|o)} [-\log(1-\hat{p}_\theta(i|x))] \; . \end{aligned} \tag{3}$$

One can derive that the Bayes optimal classifier minimizing the expected loss has the predictive distribution:

$$\hat{p}_{\theta^*}(i|x) = \frac{p(x|i)p(i)}{p(x|i)p(i) + p(x|o)p(o)} = p(i|x). \tag{4}$$

Thus at least for the training out-distribution a binary classifier based on samples from in- and (training) out-distribution would suffice to solve the OOD detection problem perfectly.

2.1. OOD detection for methods using labeled data

We now discuss how one can formulate the OOD problem in the presence of labeled in-distribution data and identify the target distribution of OOD detection using a background class. Then we derive the Bayes optimal classifier of Outlier Exposure [7] and discuss the implicit scoring function. In most cases the scoring functions turn out to be not equivalent to $p(i|x)$ (which is optimal if training and test out-distribution agree) as they integrate additional information from the classification task.

Bayes optimal solutions for OOD Detection with Background class and Outlier Exposure Given a joint in-distribution $p(y,x|i)$ (where $y \in \{1,\dots,K\}$ given that we have K labels) for the labeled in-distribution, there are different ways how to come up with a joint distribution for in- and out-distribution.

Background class: In this case we just put all out-distribution samples into a $K+1$-class which is typically called background/reject class [17]. The joint distribution then becomes

$$p(y,x) = \begin{cases} p(y,x|i)p(i) & \text{if } y \in \{1,\dots,K\}, \\ p(x|o)p(o) & \text{if } y = K+1. \end{cases}$$

We denote by $p(x|i) = \sum_{k=1}^K p(y,x|i)$ the marginal in-distribution and note that the marginal distribution of the joint distribution of in- and out-distribution is given by

$$p(x) = p(x|i)p(i) + p(x|o)p(o).$$

Then we get the conditional distribution

$$p(y|x) = \begin{cases} p(y|x,i)p(i|x) & \text{if } y \in \{1,\dots,K\}, \\ p(o|x) = 1 - p(i|x) & \text{if } y = K+1. \end{cases}.$$

The Bayes optimal solution of training with a background class using any calibrated loss function $L(y, f(x))$, e.g. the cross-entropy loss [10], yields a Bayes optimal classifier f^* which has a predictive distribution $p_{f^*}(y|x) = p(y|x)$. There are two potential scoring functions to consider:

$$s_1(x) = 1 - p_{f^*}(K+1|x), \quad s_2(x) = \max_{k=1,\dots,K} p_{f^*}(k|x)$$

The first one, used in [4, 17], is motivated by the fact that $p_{f^*}(K+1|x)$ is directly the predicted probability that the point is from the out-distribution as indeed it holds: $s_1(x) = p(i|x)$. On the other hand the maximal predicted probability $\max_{k=1,\dots,K} p_{f^*}(k|x)$, which is often employed as a scoring function [6], becomes

$$s_2(x) = p(i|x) \max_{k=1,\dots,K} p(k|x,i),$$

which is a product of $p(i|x)$ and the maximal conditional probability of some class. Thus the scoring function $s_2(x)$ additionally integrates class-specific information and is less dependent on the chosen training out-distribution. Thus s_2 assigns high rank only if both the binary discriminator *and* the classifier assign high values. However when training and test out-distribution are identical, this scoring function is not equivalent to $p(i|x)$ and introduces a bias in the estimation.

Outlier Exposure [7]: we analyze the Bayes optimal solution of Outlier Exposure (OE) and show that the associated scoring function of OE can be written, similarly to the scoring function $s_2(x)$ for training with a background class, as a function of $p(i|x)$ and $p(y|x,i)$.

The training objective of OE is in expectation given by

$$\min_{\theta} \mathop{\mathbb{E}}_{(x,y)\sim p(x,y|i)} \left[\mathcal{L}_{\text{CE}}(f_\theta(x), y)\right] + \lambda \mathop{\mathbb{E}}_{x\sim p(x|o)} \left[\mathcal{L}_{\text{CE}}(f_\theta(x), u^K)\right] \tag{5}$$

where θ are the model parameters and $f_\theta(x) \in \mathbb{R}^K$ is the logit output, and $u^K = (\frac{1}{K}, \dots, \frac{1}{K})^T$ the uniform distribution over K classes. In the following theorem we derive the Bayes optimal predictive distribution of Outlier Exposure.

Theorem 1. *The predictive distribution $p_{f^*}(y|x)$ of the Bayes optimal classifier f^* minimizing the expected OE loss is given for $y \in \{1, \dots, K\}$ as*

$$p_{f^*}(y|x) = p(i|x)p(y|x,i) + \frac{1}{K}\big(1 - p(i|x)\big) \,. \tag{6}$$

Thus the effective scoring function of OE using the probability of the predicted class is given by

$$s_3(x) = p(i|x)\Big[\max_{k=1,\dots,K} p(y|x,i) - \frac{1}{K}\Big] + \frac{1}{K}. \tag{7}$$

Please note that the term inside the brackets is positive as $\max_{k=1,\dots,K} p(k|x,i) \geq \frac{1}{K}$. Interestingly, the scoring functions s_2 and s_3 are not equivalent even though they look quite similar; due to the subtraction of $\frac{1}{K}$ the scoring function s_3 puts more emphasis on the classifier than s_2.

2.2. Shared estimation of $p(i|x)$ and $p(y|x,i)$

So far we have derived that background class OOD detection with the scoring function s_1 is equivalent to a binary classification problem between in- and out-distribution. When labeled in-distribution data is available, one can train a classifier to estimate $p(y|x,i)$. We will then combine the estimates of $p(i|x)$ and $p(y|x,i)$ according to the three scoring functions derived in the previous section and check if the novel OOD detection methods constructed in this way perform similar to the OOD methods from which we derived the corresponding scoring function i) OOD detection with a background class [17] or ii) using Outlier Exposure [7]. This will allow us to differentiate between differences of the employed scoring functions for OOD detection and the estimators for the involved quantities. In this way we foster a more systematic approach to OOD detection.

The straight forward approach for training the binary discriminator estimating $p(i|x)$ and the classifier for $p(y|x,i)$ would be to use two completely separate models that do not interact with each other during training. However, when training a neural network using a background class or with Outlier Exposure we are implicitly using a shared representation for both tasks which improves the results.

Thus we propose to train the binary discriminator of in- versus out-distribution jointly with the classifier on the in-distribution. Concretely, we use a neural network with $K+1$ outputs where the first K outputs represent the classifier and the last output is the logit of the binary discriminator. The resulting shared problem can then be written as

$$\min_{\theta} -\frac{1}{N_b}\sum_{r=1}^{N_b} \log\left(\hat{p}_\theta(i|x_r^{\text{IN}})\right) - \frac{\lambda}{M}\sum_{s=1}^{M} \log\left(1 - \hat{p}_\theta(i|x_s^{\text{OUT}})\right) - \frac{1}{N_c}\sum_{t=1}^{N_c} \log\left(\hat{p}_\theta(y_t^{\text{IN}}|x_t^{\text{IN}})\right) \tag{8}$$

where $\lambda = \frac{p(o)}{p(i)}$ which is typically set to 1 during training. We stress that the loss functions of the classifier and the discriminator act on independent outputs but share the network weights up to the final layer. In the Appendix, we have an additional evaluation where we show that the shared version significantly improves performance compared to the individual estimation of $p(i|x)$ and $p(y|x,i)$.

3. Experiments

We use the CIFAR-10 and CIFAR-100 [9] datasets as in-distribution. In our experiments, we use the 80 Million Tiny Images (80M) dataset [18] as training out-distribution as proposed by [7]. The 80M dataset is the de facto standard for training out-distribution aware models that has been adopted by most prior works. We evaluate the OOD detection performance on various out-distribution datasets which are not available during training: SVHN [14], LSUN Classroom [19], Uniform Noise, Smooth Noise generated as described by [5], the respective other CIFAR dataset, OpenImages [8], and CelebA [11]. The training of the background class model (BGC) and the shared binary discriminator+classifier (SHARED) follows the training schedule used in [7]. For OE, NTOM and ATOM, we use the models from the official repositories, see the Appendix for details.

In Table 1 we compare multiple OOD detection methods trained with training out-distribution 80M and CIFAR-10/100 as in-distribution: pre-trained models from NTOM, ATOM and OE versus a classifier with background class (BGC) and the combination of a plain classifier and a binary in-vs-out-distribution classifier with shared representation (SHARED COMBI). As described in Section 2, both BGC and SHARED COMBI can be used in combination with different scoring functions. For SHARED COMBI we only use s_2 and s_3 as s_1 is equivalent to $p(i|x)$ which is the output of SHARED BINDISC. For CIFAR-10, an interesting observation is that SHARED CLASSI has remarkably good OOD performance even though it is just trained using normal cross-entropy loss and the OOD performance is only due to the regularization enforced by the shared representation. In fact SHARED BINDISC has already very good OOD performance with a mean AUC of 97.90 which is only improved by considering scoring function s_2/s_3 in the combination of SHARED BINDISC and SHARED CLASSI which yields the

Table 1. In-distribution accuracy and AUROC against various test out-distributions for different OOD methods trained on 80M. Our binary discriminator (BINDISC) resp. the combination with the shared classifier (SHARED COMBI) performs similar/better than OE [7].

in-distribution: CIFAR-10

Model	Acc.	Mean AUC	SVHN AUC	LSUN AUC	Uni AUC	Smooth AUC	C-100 AUC	OpenIm AUC	CelA AUC	80M AUC
Plain Classi	94.84	85.68	91.91	91.63	87.85	77.92	87.83	83.23	79.43	88.06
NTOM [4]	95.42	97.32	99.59	99.79	99.97	99.84	92.19	89.96	99.89	98.72
ATOM [4]	95.20	97.42	99.63	99.76	99.93	99.60	92.89	90.30	99.85	98.55
OE [7]	95.74	97.64	99.48	99.48	99.46	99.63	94.80	90.91	99.71	98.57
BGC s_1		97.94	99.64	99.58	99.96	99.98	94.84	91.65	99.92	98.78
BGC s_2	95.63	97.95	99.60	99.52	99.97	99.98	95.03	91.65	99.92	98.71
BGC s_3	95.63	97.95	99.58	99.52	99.97	99.98	95.04	91.64	99.92	98.70
Shared BinDisc		97.90	99.74	99.60	99.94	99.96	94.75	91.42	99.87	98.82
Shared Classi	96.08	96.57	98.77	97.78	99.88	99.68	93.40	88.70	97.80	96.55
Shared Combi s_2	96.08	97.96	99.73	99.58	99.95	99.96	95.13	91.48	99.85	98.79
Shared Combi s_3	96.08	97.96	99.73	99.58	99.96	99.96	95.14	91.48	99.85	98.78

in-distribution: CIFAR-100

Model	Acc.	Mean AUC	SVHN AUC	LSUN AUC	Uni AUC	Smooth AUC	C-10 AUC	OpenIm AUC	80M AUC
Plain Classi	75.96	77.45	71.38	76.89	78.13	88.35	75.33	74.60	76.14
NTOM [4]	74.88	88.49	96.20	97.31	99.79	99.94	62.44	75.24	88.41
ATOM [4]	75.06	88.02	93.68	97.51	99.98	98.46	63.47	75.02	88.44
OE [7]	76.73	91.72	94.06	95.58	99.05	98.80	79.53	83.31	88.96
BGC s_1		92.04	94.42	95.47	99.99	99.72	79.15	83.46	89.72
BGC s_2	75.82	91.54	93.32	94.64	99.95	99.61	79.29	82.41	88.77
BGC s_3	75.82	91.53	93.30	94.62	99.95	99.61	79.29	82.40	88.76
Shared BinDisc		91.84	95.90	95.69	99.79	99.95	76.56	83.19	91.91
Shared Classi	76.52	88.16	86.28	87.61	99.97	99.91	77.00	78.23	85.33
Shared Combi s_2	76.52	92.03	95.50	95.43	99.96	99.98	78.46	82.86	91.44
Shared Combi s_3	76.52	92.03	95.49	95.42	99.97	99.98	78.46	82.85	91.43

best performance. Moreover, the classifier with background class (BGC) can work very well with slight but noticeable differences between the scoring functions. The background class models reach SOTA performance similar to/better than OE (which implicitly also uses s_3). However, overall the differences of the methods are relatively minor.

The results for CIFAR-100 are similar to CIFAR-10. NTOM/ATOM now show significantly worse OOD results which are mainly due to the close out-distributions CIFAR-10 and OpenImages. OE achieves comparable OOD results to the results of BGC and SHARED COMBI s_2/s_3. In particular, SHARED CLASSI on its own is able to perform similar to NTOM and ATOM without being explicitly trained for OOD detection. Our BGC s_1 and SHARED COMBI s_2/s_3 work best in terms of OOD performance and show good test accuracy, but again differences are minor.

Overall, as suggested by the theoretical results on the equivalence of the scoring functions associated to Bayes optimal classifier of OE with its implicit scoring function s_3, BGC with s_3 and SHARED COMBI with s_3, we observe that despite these methods being implemented quite differently, they behave very similar in our experimental results as suggested by the theoretical considerations. In total we think that this provides a much better understanding where differences of OOD methods are coming from, and shows that for potential applications, one has some freedom of choice between the different methods. Moreover, we have shown that our proposed SHARED COMBI achieves arguably the best combination of mean AUC and test accuracy for both datasets.

4. Conclusion

In this paper we have analyzed different OOD detection methods and have shown that the simple baseline of a binary discriminator between in-and out-distribution is already a very powerful OOD detection method if trained in a shared fashion with a classifier. Moreover, we have revealed the inner mechanism of Outlier Exposure and training with a background class which combine information from $p(i|x)$ and $p(y|x, i)$.

References

[1] M. Augustin, A. Meinke, and M. Hein. Adversarial robustness on in-and out-distribution improves explainability. In *ECCV*, 2020.

[2] A. Birhane and V. U. Prabhu. Large image datasets: A pyrrhic win for computer vision? In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1537–1547, January 2021.

[3] S. Boyd, S. P. Boyd, and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[4] J. Chen, Y. Li, X. Wu, Y. Liang, and S. Jha. Informative outlier matters: Robustifying out-of-distribution detection using outlier mining. *arXiv preprint arXiv:2006.15207*, 2020.

[5] M. Hein, M. Andriushchenko, and J. Bitterwolf. Why ReLU networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In *CVPR*, 2019.

[6] D. Hendrycks and K. Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *ICLR*, 2017.

[7] D. Hendrycks, M. Mazeika, and T. Dietterich. Deep anomaly detection with outlier exposure. In *ICLR*, 2019.

[8] I. Krasin, T. Duerig, N. Alldrin, V. Ferrari, S. Abu-El-Haija, A. Kuznetsova, H. Rom, J. Uijlings, S. Popov, S. Kamali, M. Malloci, J. Pont-Tuset, A. Veit, S. Belongie, V. Gomes, A. Gupta, C. Sun, G. Chechik, D. Cai, Z. Feng, D. Narayanan, and K. Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://storage.googleapis.com/openimages/web/index.html*, 2017.

[9] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

[10] D. Laptev, N. Savinov, J. Buhmann, and M. Pollefeys. TI-pooling: Transformation-invariant pooling for feature learning in convolutional neural networks. In *CVPR*, 2016.

[11] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *ICCV*, 2015.

[12] A. Meinke and M. Hein. Towards neural networks that provably know when they don't know. In *ICLR*, 2020.

[13] S. Mohseni, M. Pitale, J. Yadawa, and Z. Wang. Self-supervised learning for generalizable out-of-distribution detection. In *AAAI*, 2020.

[14] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NeurIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

[15] A. Nguyen, J. Yosinski, and J. Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *CVPR*, 2015.

[16] A.-A. Papadopoulos, M. R. Rajati, N. Shaikh, and J. Wang. Outlier exposure with confidence control for out-of-distribution detection. *Neurocomputing*, 441:138–150, 2021.

[17] S. Thulasidasan, S. Thapa, S. Dhaubhadel, G. Chennupati, T. Bhattacharya, and J. Bilmes. An effective baseline for robustness to distributional shift. *arXiv: 2105.07107*, 2021.

[18] A. Torralba, R. Fergus, and W. T. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 30(11):1958–1970, 2008.

[19] F. Yu, Y. Zhang, S. Song, A. Seff, and J. Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *CoRR*, abs/1506.03365, 2015.

[20] S. Zagoruyko and N. Komodakis. Wide residual networks. In *BMVC*, pages 87.1–87.12, 2016.

A. Proof Theorem 1

Proof. This is the minimization problem

$$\begin{aligned}\min_{p_\theta(x)} \quad & -p(\text{in}\,|x) \cdot \sum_{k=1}^{K} p(x|i)(k|x) \cdot \log p_\theta(x)[k] - (1 - p(\text{in}\,|x)) \cdot \sum_{k=1}^{K} \frac{1}{K} \cdot \log p_\theta(x)[k] \\ \text{subject to} \quad & p_\theta(x)[k] \geq 0 \;\text{ for each } k \in \{1, \dots, K\} \\ & \sum_{k=1}^{K} p_\theta(x)[k] = 1\,.\end{aligned}$$

For $p(\text{in}\,|x) = 0$ or $p(x|i)(k|x) = 0$, the optimalities of the respective terms are easy to show (applying the common conventions for $0 \log 0$), so we assume those to be non-zero. The Lagrange function of the optimization problem is

$$\begin{aligned}L(p_\theta(x), \alpha, \beta) = & -p(\text{in}\,|x) \cdot \sum_{k=1}^{K} p(x|i)(k|x) \cdot \log p_\theta(x)[k] - (1 - p(\text{in}\,|x)) \cdot \sum_{k=1}^{K} \frac{1}{K} \cdot \log p_\theta(x)[k] \\ & - \sum_{k=1}^{K} \alpha_k p_\theta(x)[k] + \beta \left(-1 + \sum_{k=1}^{K} p_\theta(x)[k] \right)\,,\end{aligned}$$

with $\beta \in \mathbb{R}$ and $\alpha \in \mathbb{R}_+^K$. Its first derivative with respect to $p_\theta(x)[k]$ is

$$\begin{aligned}\frac{\partial L}{p_\theta(x)[k]} &= -p(\text{in}\,|x) \cdot p(x|i)(k|x) \frac{1}{p_\theta(x)[k]} - (1 - p(\text{in}\,|x)) \cdot \frac{1}{K} \frac{1}{p_\theta(x)[k]} - \alpha_k + \beta \\ &= -\frac{s^K(x)[k]}{p_\theta(x)[k]} - \alpha_k + \beta.\end{aligned} \tag{9}$$

The second derivative is a positive diagonal matrix on the domain, therefore we find the unique minimum by setting 9 to zero, which means

$$p_\theta(x)[k] = \frac{s^K(x)[k]}{\beta - \alpha_k}\,.$$

The dual problem is hence maximizing (with $\alpha_k \geq 0$)

$$\begin{aligned}q(\alpha, \beta) = & -p(\text{in}\,|x) \cdot \sum_{k=1}^{K} p(x|i)(k|x) \cdot \log \frac{s^K(x)[k]}{\beta - \alpha_k} - (1 - p(\text{in}\,|x)) \cdot \sum_{k=1}^{K} \frac{1}{K} \cdot \log \frac{s^K(x)[k]}{\beta - \alpha_k} \\ & - \sum_{k=1}^{K} \alpha_k \frac{s^K(x)[k]}{\beta - \alpha_k} + \beta \left(-1 + \sum_{k=1}^{K} \frac{s^K(x)[k]}{\beta - \alpha_k} \right) \\ = & \sum_{k=1}^{K} s^K(x)[k] \left(-\log s^K(x)[k] + \log(\beta - \alpha_k) + \frac{\beta}{\beta - \alpha_k} - \frac{\alpha_k}{\beta - \alpha_k} \right) - \beta\,;\end{aligned}$$

here, α only appears in $\log(\beta - \alpha_k)$, so $\alpha = 0$ maximizes the expression. Noting $\sum_{k=1}^{K} s^K(x)[k] = 1$, what remains is $q^0(\beta) = 1 + \log(\beta) - \sum_{k=1}^{K} s^K(x)[k] \log s^K(x)[k] - \beta$, which is maximized by $\beta = 1$. This means that the dual optimal pair is $p_\theta(x)[k] = s^K(x)[k], (\beta = 1, \alpha = 0)$. Slater's condition [3] holds since the feasible set of the original problem is the probability simplex. Thus, $p_\theta(x) = s^K(x)$ is indeed primal optimal. □

B. Experimental details

The binary discriminators (BINDISC) as well as the classifiers with background class (BGC) and the shared binary discriminator+classifier (SHARED) use the Wide ResNet 40-2 [20] architecture with the same code and training schedule as used in [7]. This way we ensure that the differences do not arise due to differences in the training schedules. For the established methods, we use the Plain and Outlier Exposure (OE) models from the official OE codebase[1] and for NTOM and ATOM we use their code [2] to evaluate their DenseNet models (their best models).

In our experiments, we use the 80 Million Tiny Images (80M) dataset [18] as training out-distribution as proposed by [7]. The 80M dataset is the de facto standard for training out-distribution aware models that has been adopted by most prior works. Recently, this dataset has been retracted by the authors after [2] had pointed out the presence of offensive and prejudicial images. We still use this dataset to be able to compare with other state-of-the-art methods without introducing a potential bias due to dataset selection.

C. Shared versus individual estimation of $p(i|x)$ and $p(y|x,i)$

As highlighted in the main experimental results section, the shared training of SHARED CLASSI and SHARED BINDISC and their combination SHARED COMBI with s_2/s_3 as scoring functions yields the best OOD detection and test accuracy among all methods. In this section, we briefly evaluate the importance of training the binary discriminator and the plain classifier with a shared representation in comparison to training two entirely separate models PLAIN CLASSI and SEPARATE BINDISC and their combination SEPARATE COMBI with scoring function s_3. The results for CIFAR-10 and CIFAR-100 can be found in Table 2. In total, we see that separate training in particular for CIFAR-100 leads to significantly worse results compared to shared training as expected as the binary discriminator and the classifier cannot benefit from each other. An interesting curiosity is that the combination of the separate classifier with the binary discriminator trained in a shared fashion (PLAIN $\otimes$ SHA DISC) yields almost the same OOD results as SHARED COMBI even though the classifier is significantly worse. Overall, SHARED COMBI performs significantly better when also considering the better classification accuracy which it inherits of SHARED CLASSI.

Table 2. Evaluation (same metrics as in Table 1) of models trained with shared and separate representations. Shared training benefits both the classifier and the binary discriminators.

IN-DISTRIBUTION: CIFAR-10

MODEL	ACC.	MEAN	SVHN	LSUN	UNI	SMOOTH	C-100	OPENIM	CELA	80M
PLAIN CLASSI	94.84	85.75	91.91	91.63	87.69	78.27	87.83	83.23	79.43	88.42
SEPARATE BINDISC		97.06	99.53	99.59	99.94	99.99	90.59	89.93	99.84	98.86
SEPARATE COMBI s_3	94.84	97.50	99.53	99.54	99.93	99.99	93.10	90.56	99.84	98.94
SHARED BINDISC		97.90	99.74	99.60	99.94	99.96	94.75	91.42	99.87	98.82
SHARED CLASSI	96.08	96.57	98.77	97.78	99.88	99.68	93.40	88.70	97.80	96.55
SHARED COMBI s_3	96.08	97.96	99.73	99.58	99.96	99.96	95.14	91.48	99.85	98.78
PLAIN $\otimes$ SHA DISC s_3	94.84	97.95	99.72	99.56	99.93	99.96	95.18	91.47	99.85	98.60

IN-DISTRIBUTION: CIFAR-100

MODEL	ACC.	MEAN	SVHN	LSUN	UNI	SMOOTH	C-10	OPENIM	80M
PLAIN CLASSI	75.96	77.48	71.38	76.89	78.14	88.36	75.33	74.60	76.24
SEPARATE BINDISC		78.28	54.41	94.84	99.95	97.99	54.74	67.77	82.56
SEPARATE COMBI s_3	75.96	82.58	61.75	95.02	99.94	98.76	66.01	74.00	85.71
SHARED BINDISC		91.84	95.90	95.69	99.79	99.95	76.56	83.19	91.91
SHARED CLASSI	76.52	88.16	86.28	87.61	99.97	99.91	77.00	78.23	85.33
SHARED COMBI s_3	76.52	92.03	95.49	95.42	99.97	99.98	78.46	82.85	91.43
PLAIN $\otimes$ SHA DISC s_3	75.96	91.89	95.23	95.19	99.71	99.94	78.49	82.78	88.70

[1] https://github.com/hendrycks/outlier-exposure
[2] https://github.com/jfc43/informative-outlier-mining