When more is less
Adverse Effects in Outlier Exposure

Jennifer Jorina Matthiesen* and Ulf Brefeld

Machine Learning Group, Leuphana University of Lüneburg
{jennifer.matthiesen,brefeld}@leuphana.de

1 Introduction

Anomaly detection (AD) [1, 2] is the task of identifying items that differ from the majority of data. While AD is addressing the distinction between normal and abnormal data, it is typically treated as an unsupervised learning problem, since it is rather difficult to find or construct a counter class capturing *everything else* outside the regarded target class.

As in many AD methods [8, 6], the underlying objective is to find the smallest enclosure of data with minimum volume by using a hyperplane or a hypersphere respectively. However, based on their objective, those methods suffer from a hypersphere collapse if the architecture of the model does not comply with certain constraints. Recently, outlier exposure (OE) [3] aims to overcome the sparsity of data by including auxiliary class labels in the training of the anomaly detector for robustness.

Using OE data during training promises stunning results [4, 7], while just using a number of 128 random samples in the OE dataset. However, we show that the selection of those samples actually has a major influence on the detector's performance. Intuitively, one might think that when learning features for a target class, including a variety of alternative classes (anomalies) should increase performance. We show that this does not always hold. Using more classes in the OE dataset during training can likewise result in decreasing the performance. We study multiple variations of training data and its effects on the performance and observe significant dependencies of the target data and data used as OE that may either foster or prohibit predictive performance. We show that (i) random samples of auxiliary data may not give optimal results and (ii) while increasing the number of classes in the auxiliary OE data (i.e. increasing the variety of anomalies), can in fact result in decreased performance.

2 Network Architecture

Let $D = \{(x_1, y), (x_2, y), \ldots, (x_n, y)\}$ be a dataset with $x_i \in \mathbb{R}^d$ and $y \in \{0, 1, \ldots, m\}$. Here D is a composition of a target dataset D_T and an auxiliary dataset D_{OE}, where $y = 0$ denotes the nominal data, while $y \in \{1, 2, \ldots, m\}$ denotes anomalies. Let $\phi_f : \mathbb{R}^d \to \mathbb{R}^r$ be a neural network for feature extraction of D_{T} and D_{OE} and $\phi_q : \mathbb{R}^r \to \{1, 2, \ldots, m\}$ a sub network for classification of the features produces by ϕ_f.

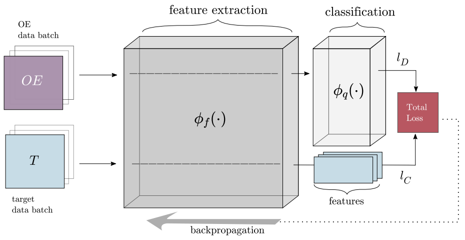


Figure 1: Networks' architecture used in this work. The network $\phi_f(\cdot)$ extracts features from both datasets, while additional layers of $\phi_q(\cdot)$ are used for the auxiliary data to calculate l_D.

Using additional data during training, we propose a composite loss $l_D(D_{OE}) + \lambda l_C(D_T)$ to address the two desired characteristics comparativeness and descriptiveness (cmp Patel [5]), given by a compactness loss l_C computed by squared intra-

*Corresp. Author: jennifer.matthiesen@leuphana.de batch distances

$$l_C = \frac{1}{n}\sum_{i=1}^{n}(\phi_f(x_i;\theta) - m_i)^2, \qquad (1)$$

and a descriptive loss l_D being the cross entropy loss over all classes C,

$$l_D = -\sum_{i=1}^{C} y_i log(\phi_q(x_i;\theta))^2 \qquad (2)$$

where $m_i = \frac{1}{n-1}\sum_{j\neq i}(\phi_f(x_j;\theta))$ is the mean vector of the rest of the features of the regarded sample. We make use of the AlexNet CNN architecture for the structure of ϕ_f.

3 Experiments

Using a multi-class dataset like CIFAR, we follow the common one vs. rest setting for anomaly detection. All models are trained on the target class D_T as well as on OE classes D_{OE} , while test data includes the test set of the target class and the remaining classes, treated as anomalies D_A, of the considered dataset. We study how the auxiliary dataset D_{OE} should be assembled given a target class D_T. We use CIFAR-10 classes as normal and abnormal data, while selected classes of CIFAR-100 are used as outlier data D_{OE}.

To determent the influence of each separate class, we run further testing using each time one of the nine classes. In contrast to related literature, we are not using more classes in OE, but try to gain a better understanding of the usage of OE in AD through the reduction of factors. Note that, reducing the number of influence factors also results in excluding a pre-trained network, since the pretrained layers include already a certain influence from the network they are trained with.

3.1 Results

The results in Table 2 show counter-intuitive as well as non-surprising outcomes. For example, including whales in D_{OE} constantly deteriorates performance. Distinguishing between dogs and any other class improves when wolfs are present in D_{OE}. The same result holds true for cats in D_{OE}. If both D_T and D_A are animals, it is beneficial to have cattle in D_{OE}.

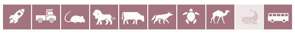
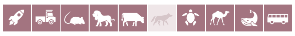
	Training with/without whales				
Cls. in training:	D_{OE}				
Cls. in testing:	D_T	D_A	AUC	AUC$_{\text{all}}$	diff.
	plane	rest	82.1	**79.6**	**0.25**
	bird	rest	73.7	**71.3**	**0.24**
	deer	rest	75.9	**73.3**	**0.26**
	cat	rest	73.1	74.2	-0.11
	Training with/without wolves				
Cls. in training:	D_{OE}				
Cls. in testing:	D_T	D_A	AUC	AUC$_{\text{all}}$	diff.
	plane	dog	83.9	87.1	-0.32
	bird	dog	60.5	60.6	-0.05
	deer	dog	68.1	72.2	-0.41
	cat	dog	54.3	54.9	-0.06

Figure 2: Leaving out whales, results in a better classifier for most of the target classes, while wolves are beneficial when dogs are in D_A.

4 Conclusions

We showed that using the selection of auxiliary classes in OE is crucial for the performance of the classifier. To the best of our knowledge, this effect has not been studied in the literature so far. We demonstrated that increasing numbers of classes in D_{OE} do not *per se* result in better performances. Even worse, putting together the wrong classes will significantly decrease the detection accuracy. Note that an OE sample from more classes might compensate for the effect of some classes. However, making a careful selection can help to increase the performance. This study constituted a first step in determining the influence of classes in OE, which can help to choose additional classes for training in AD problems. However, more experimentation is necessary.

References

[1] C. C. Aggarwal. *Outlier Analysis*. Springer Publishing Company, Incorporated, 2nd edition, 2016.

[2] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):15, 2009.

[3] D. Hendrycks, M. Mazeika, and T. Dietterich. Deep anomaly detection with outlier exposure.

In *International Conference on Learning Representations*, 2019.

[4] D. Hendrycks, M. Mazeika, S. Kadavath, and D. Song. Using self-supervised learning can improve model robustness and uncertainty. In H. Wallach, H. Larochelle, A. Beygelzimer, F. Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[5] P. Perera and V. M. Patel. Learning deep features for one-class classification. *IEEE Transactions on Image Processing*, 28(11):5450–5463, 2019.

[6] L. Ruff, R. Vandermeulen, N. Goernitz, L. Deecke, S. A. Siddiqui, A. Binder, E. Müller, and M. Kloft. Deep one-class classification. In *International Conference on Machine Learning*, pages 4393–4402, 2018.

[7] L. Ruff, R. A. Vandermeulen, B. J. Franks, K.-R. Müller, and M. Kloft. Rethinking assumptions in deep anomaly detection, 2020.

[8] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the Support of a High-Dimensional Distribution. *Neural computation*, 13(7):1443–1471, 2001.