

Mahalanobis-Aware Training for Out-of-Distribution Detection

Connor Mclaughlin*, Jason Matterer†, and Michael Yee

MIT Lincoln Laboratory
mclaughlin.co@northeastern.edu, jason.matterer@str.us, myee@ll.mit.edu

Abstract

While deep learning models have seen widespread success in controlled environments, there are still barriers to their adoption in open-world settings. One critical task for safe deployment is the detection of anomalous or out-of-distribution samples that may require human intervention. In this work, we present a novel loss function and recipe for training networks with improved density-based out-of-distribution sensitivity. We demonstrate the effectiveness of our method on CIFAR-10, notably reducing the false-positive rate of the relative Mahalanobis distance method on far-OOD tasks by over 50%.

Introduction

Deep neural networks have achieved widespread ubiquity across decision-making processes in various applications, ranging from medical imaging to autonomous driving. When deployed to the open world, these systems frequently encounter samples from classes not represented in the training data, known as “out-of-distribution” (OOD) data. In these settings, it is essential for the model to have the ability to abstain or hand off the difficult decision to a human expert. Consequently, the development of effective OOD detection techniques is crucial for enhancing the robustness and reliability of these systems in real-world scenarios.

Existing OOD detection methods typically fall under two categories: (1) model output-based methods, which take into account the predicted logits or probabilities of the model as an indicator of confidence, or (2) model representation-based methods, which measure the similarity of intermediate layer representations to those seen at training time. Our study focuses on the latter class of methods, which rely on learning a representative model of the in-distribution (ID) data in order to compute the likelihood of test samples. Some studies assume a Gaussian structure for the data (Lee et al. 2018), achieving empirical success but leaving much of the theory unexplored. In contrast, others adopt more complex, non-parametric methods (Sun et al. 2022) to avoid imposing any assumptions on the data. In this paper, we aim to

*Work done at MIT LL; now at Northeastern University

†Work done at MIT LL; now at STR

Method	Far-OOD		Near-OOD	
	FPR95 ↓	AUROC ↑	FPR95 ↓	AUROC ↑
MSP	50.91	90.62	61.25	89.67
Energy	40.97	89.10	51.30	86.24
KNN	37.10	94.47	52.31	89.67
ViM	27.50	95.66	55.25	87.66
RMD	36.58	93.80	51.54	89.01
Ours	17.79	96.76	49.53	90.43

Table 1: Results using CIFAR-10 as in-distribution dataset.

contribute to this discourse by investigating the following question: Can we improve the performance of Gaussian-based OOD detection methods through explicit model training aimed at creating Gaussian-like data representations?

Our contributions are twofold: (1) we first present a novel regularization loss that better aligns the training-time objective and test-time OOD detector, and (2) we also provide a training recipe with noise reduction techniques that make our method more accessible to a limited computational budget. Our proposed method results in a significant leap in sensitivity to OOD instances, all while maintaining minimal disruption to in-distribution performance.

Methods

The task of OOD detection hinges on learning a scoring function $S(\mathbf{x})$ which captures the similarity of test data to the training distribution. In the case of density estimation methods, this scoring function is akin to the likelihood function of a probabilistic model representing the in-distribution data. Combined with a threshold τ , the OOD decision rule can be formalized as follows:

$$\text{Decision}(\mathbf{x}) = \begin{cases} \text{ID}, & \text{if } S(\mathbf{x}) \geq \tau \\ \text{OOD}, & \text{if } S(\mathbf{x}) < \tau \end{cases}$$

The Mahalanobis-Distance method (Lee et al. 2018) models the distribution of latent representations of the neural network as coming from a class-conditional Gaussian with means $\mu_{1..k}$ and a tied covariance matrix Σ . The most commonly selected representation is the output of the penultimate layer of the network, denoted as $\mathbf{z} = F(\mathbf{x})$ for input \mathbf{x} .

The scoring function is given as the Mahalanobis distance of representation \mathbf{z} to the closest in-distribution class centroid (with flipped sign so ID samples have higher score):

$$\text{MD}_k(\mathbf{z}) = (\mathbf{z} - \boldsymbol{\mu}_k)^T \boldsymbol{\Sigma}^{-1} (\mathbf{z} - \boldsymbol{\mu}_k)$$

$$S_{\text{MD}}(\mathbf{z}) = -\min_k \text{MD}_k(\mathbf{z})$$

As the success of Mahalanobis distance relies upon ID samples having higher likelihoods than OOD samples, our key insight is in promoting the likelihood of ID data throughout training. Our proposed training objective utilizes an on-line estimate of Gaussian parameters in order to compute the predicted probability of test samples under Bayes' rule:

$$\begin{aligned} P(Y = k|X = x) &= \frac{P(Y = k)P(X = x|Y = k)}{\sum_{k'} P(Y = k')P(X = x|Y = k')} \\ &= \frac{\exp(\text{MD}_k)}{\exp(\sum_{k'} \text{MD}_{k'})} \end{aligned}$$

We then look to minimize the cross-entropy loss using these predictions. Our final combined loss is a weighted combination of the initial cross-entropy loss using model output logits (L_{base}) and the cross-entropy loss using Mahalanobis distances as logits (L_{maha}). We introduce a hyperparameter α to control the balance between losses:

$$L_{\text{reg}} = (1 - \alpha)L_{\text{base}} + \alpha L_{\text{maha}}$$

The final component of our proposed method is in estimating the Gaussian parameters throughout training. As the batch size n may be much smaller than the dimensionality of our feature representations d , we introduce two necessary components to reduce the noise introduced by small batch estimates. First, we use a shrinkage estimator for the covariance matrix (Ledoit and Wolf 2004) rather than the maximum likelihood estimator. Second, we maintain a moving average (EMA) of the means and covariance which is updated with each batch of training data.

Once the training is complete, our method utilizes the Relative Mahalanobis (RMD) score (Ren et al. 2021) variation of Mahalanobis distance as the OOD scoring function.

Results

We demonstrate the efficacy of our training policy through a series of experiments using CIFAR-10 as the in-distribution dataset. We use the ResNet18 architecture and follow the baseline CIFAR-10 setup provided by OpenOOD (Yang et al. 2022). Table 1 compares our method to recent approaches on a far-OOD benchmark consisting of (SVHN, Places365, iSUN, LSUN, and Textures), and a more challenging near-OOD benchmark consisting of CIFAR-100. We compare our method to MSP (Hendrycks and Gimpel 2016), Energy (Liu et al. 2020), RMD (Ren et al. 2021), KNN (Sun et al. 2022), and ViM (Wang et al. 2022). Our method outperforms existing baselines in both near-OOD and far-OOD settings, making it a reliable choice regardless of the deployment environment. We additionally show that our Mahalanobis loss has the intended effect on the Gaussian likelihood of data representations in Figure 1. Our method is not

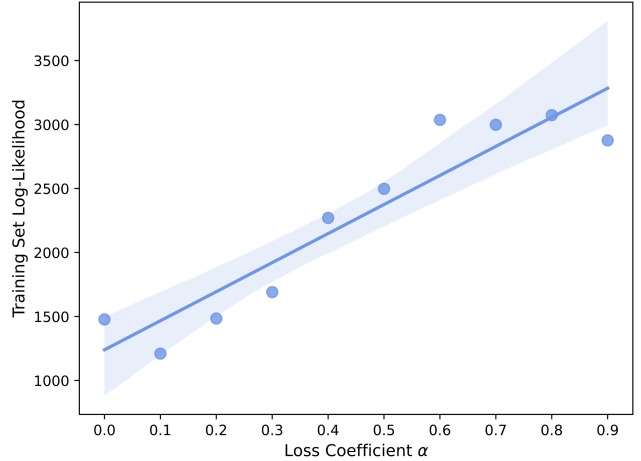


Figure 1: Training Dataset Gaussian LL vs. Loss α

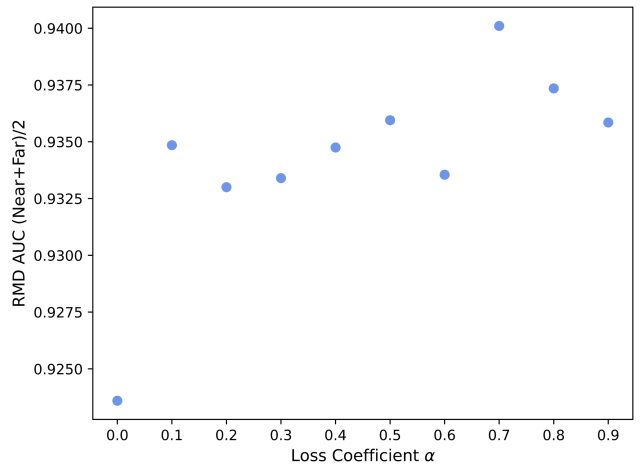


Figure 2: Sensitivity Analysis of Parameter α

sensitive to the added hyperparameter α , and consistently outperforms the baseline model as shown in Figure 2. In-distribution performance remains steady, with our method achieving 94.7% accuracy compared to the 95.2% accuracy of the baseline recipe.

Conclusion

In summary, we present a novel training recipe based on Mahalanobis distance regularization for improved out-of-distribution detection. Future work includes exploring the scalability of this method to large scale datasets.

Acknowledgements

DISTRIBUTION STATEMENT A. Approved for public release. Distribution is unlimited.

This material is based upon work supported by the Under Secretary of Defense for Research and Engineering under Air Force Contract No. FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this

material are those of the author(s) and do not necessarily reflect the views of the Under Secretary of Defense for Research and Engineering.

References

- Hendrycks, D.; and Gimpel, K. 2016. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*.
- Ledoit, O.; and Wolf, M. 2004. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2): 365–411.
- Lee, K.; Lee, K.; Lee, H.; and Shin, J. 2018. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31.
- Liu, W.; Wang, X.; Owens, J.; and Li, Y. 2020. Energy-based out-of-distribution detection. *Advances in neural information processing systems*, 33: 21464–21475.
- Ren, J.; Fort, S.; Liu, J.; Roy, A. G.; Padhy, S.; and Lakshminarayanan, B. 2021. A simple fix to mahalanobis distance for improving near-ood detection. *arXiv preprint arXiv:2106.09022*.
- Sun, Y.; Ming, Y.; Zhu, X.; and Li, Y. 2022. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, 20827–20840. PMLR.
- Wang, H.; Li, Z.; Feng, L.; and Zhang, W. 2022. Vim: Out-of-distribution with virtual-logit matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4921–4930.
- Yang, J.; Wang, P.; Zou, D.; Zhou, Z.; Ding, K.; Peng, W.; Wang, H.; Chen, G.; Li, B.; Sun, Y.; et al. 2022. Openood: Benchmarking generalized out-of-distribution detection. *Advances in Neural Information Processing Systems*, 35: 32598–32611.