Falsehoods that ML researchers believe about OOD detection

Andi Zhang [1] Damon Wischik [1]

[1]Department of Computer Science and Technologies, University of Cambridge

A paradoxical result

An intuitive way to detect out-of-distribution (OOD) data is via the density function of a fitted probabilistic generative model: points with low density may be classed as OOD. But this naive approach leads to a paradoxical result, as shown by Nalisnick et al. [6]:

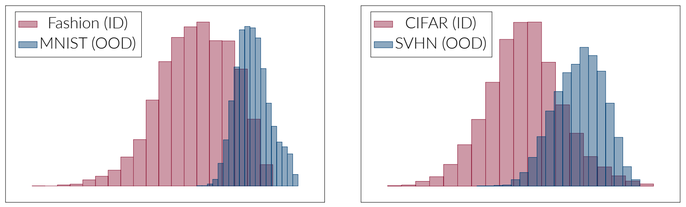

Figure 1. The left plot shows a PixelCNN model that is trained on FashionMNIST and tested on FashionMNIST (ID) and MNIST (OOD); the right plot show a PixelCNN model that is trained on CIFAR10 and tested on CIFAR10 (ID) and SVHN (OOD). The x-axis indicates the log-likelihood normalised by the data dimension and y-axis represents the data counts. We can observe that OOD datasets consistently obtain higher test likelihood than ID datasets. Plots are derived from [14].

Falsehoods

In fact, the naive approach to OOD detection is based on several falsehoods:

- that $p(x)$ should be lower on OOD data;
- that the paradoxical result arises from some deep-learning dark magic;
- that $p(x)$ is suitable for comparing two distributions;
- that low $p(x)$ indicates lack of samples.

It is not a paradox

Suppose the training dataset is drawn from $N(0,1)$, and that the training procedure has correctly learned the density $p(x) = \mathcal{N}(x;0,1)$. Now consider an OOD dataset drawn from $N(0,\varepsilon^2)$ for some small ε. Then the expected log likelihoods are

$$\mathbb{E}\log p(X) = \frac{1}{2}\log 2\pi - \begin{cases} 1/2 & \text{for in-distribution i.e. } X \sim N(0,1) \\ \varepsilon^2/2 & \text{for OOD i.e. } X \sim N(0,\varepsilon^2). \end{cases}$$

We see that $\log p(X)$ is larger for out-of-distribution data. This isn't a paradox, it's expected behaviour! And it arises from basic probability, not from mysterious properties of deep generative modelling.

Nalisnick et al. [6]: 'Our conclusion is that SVHN simply "sits inside of" CIFAR10—roughly same mean, smaller variance—resulting in its higher likelihood.'

Lack of samples?

Why was the result of Nalisnick et al. [6] surprising? The intuition is something like this: the training dataset (CIFAR10) has no samples that look anything like the OOD dataset (SVHN), therefore we expect $p(x)$ to be low on those OOD datapoints.

But, "Gaussian distributions are soap bubbles" [2]. The pdf is always highest at the origin, and yet we are very unlikely to see any sample points in a ball around the origin! [7] In other words, "lack of samples" should not be confused with "low pdf".

Comparing two distributions

Bishop [1] pointed out that OOD detection can be thought of as model selection between the in-distribution p_{in} and an out-of-distribution p_{out}.

In the paper, we show that **the likelihood ratio is an optimal choice from both frequentist and Bayesian perspectives.** However, it is hard to obtain p_{out}.

OOD proxies

Several recent works on OOD detection can however be thought of as using a likelihood ratio test based on a *proxy* distribution for p_{out}. Formally, we can propose an OOD proxy $p_{\text{out}}^{\text{proxy}}$, and use the likelihood ratio $p_{\text{in}}/p_{\text{out}}^{\text{proxy}}$ as our OOD criterion.

Constant. Bishop [1] suggested we take $p_{\text{out}}^{\text{proxy}}$ to be a constant because he thought that p_{out} should spread out widely in a large area. Then the likelihood ratio is identical to $p(x)$ used by Nalisnick et al. [6]. They reported that this choice of $p_{\text{out}}^{\text{proxy}}$ leads to poor performance, as measured by AUROC, in deep learning examples.

Auxiliary OOD datasets. It is natural to construct $p_{\text{out}}^{\text{proxy}}$ by some real out-of-distribution data. Hendrycks et al. [5] suggested that introducing an auxiliary OOD data (e.g. 80 Million Tiny Images [3]) will increase the anomaly detection performance. Then, Schirrmeister et al. [11] proposed a criterion using likelihood ratio between in-distribution p_{in} and general image distribution p_{g}, where p_{g} is trained by the aforementioned auxiliary OOD dataset, i.e. the $p_{\text{out}}^{\text{proxy}}$. Furthermore, Zhang et al. [15] suggested that the likelihood ratio could be estimated by a binary classifier.

Local features. Zhang et al. [14] proposed detecting OOD by using local models, i.e. models constrained to capture only limited perceptual fields of the image. They observed that the local models and full models assign similar likelihoods to OOD data, and infer that the local features are shared between in-distribution and OOD datasets while non-local features are not. They assume that the full model admits a decomposition $p_{\text{in}}(x) \propto p_{\text{in}}^{\text{local}}(x)\, p_{\text{in}}^{\text{nonlocal}}(x)$, and propose that $p_{\text{in}}^{\text{nonlocal}}$ should be used for detecting OOD data. This can be written as

$$p_{\text{in}}^{\text{nonlocal}}(x) \propto \frac{p_{\text{in}}(x)}{p_{\text{in}}^{\text{local}}(x)}$$

which is our general-purpose likelihood ratio criterion, using the local model trained on in-distribution data as the proxy OOD distribution.

Likelihood-ratio is not a hack

Most of the works introduced in **OOD proxies** use 'failure' or some similar words to describe the phenomenon reported by Nalisnick et al. [6]. They proposed solutions or patches based on background statistics, local features, or data complexity to "fix the issue"; and all of them have a final form in likelihood ratio. According to Bishop [1], and as we discussed in **Comparing two distributions**, density-based OOD detection is a special case of likelihood-ratio-based OOD detection. Hence, we emphasise that likelihood ratio is not a hack to fix density-based detection, it is a principled way to detect OOD.

Discussion

Semantics *v.* domain distinction. The works we have discussed [9, 11, 12, 14, 15] include interpretations in the language of semantics. Indeed, the benchmark proposed by Hendrycks and Gimpel [4] is semantic: "We can see that SVHN is semantically different to CIFAR10, so SVHN should be considered OOD." But it's hard to define 'semantics' rigorously, and so semantic-based OOD detection can seem *ad hoc*. In our opinion, it's simpler to treat OOD detection as just a problem of detecting domains (p_{in} versus $p_{\text{out}}^{\text{proxy}}$), and this leads directly to the very clean answer "use likelihood ratio".

Generalisation of OOD proxies. We want an OOD proxy that can distinguish the in-distribution domain from other domains, which is so-called generalization. Hendrycks et al. [5] indicated that using real auxiliary data as the OOD proxy has a better performance than using the augmented in-distribution data. Investigating the generalisation of OOD proxies is an open question, which we leave to future work.

References

[1] Christopher M Bishop. Novelty detection and neural network validation. *IEE Proceedings-Vision, Image and Signal processing*, 141(4):217–222, 1994.

[2] Avrim Blum, John Hopcroft, and Ravindran Kannan. *Foundations of data science*. Cambridge University Press, 2020.

[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.

[4] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

[5] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018.

[6] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know? *International Conference on Learning Representations*, 2019.

[7] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, and Balaji Lakshminarayanan. Detecting out-of-distribution inputs to deep generative models using a test for typicality. *arXiv preprint arXiv:1906.02994*, 5, 2019.

[8] Jerzy Neyman and Egon Sharpe Pearson. Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A. Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, 1933.

[9] Jie Ren, Peter J Liu, Emily Fertig, Jasper Snoek, Ryan Poplin, Mark Depristo, Joshua Dillon, and Balaji Lakshminarayanan. Likelihood ratios for out-of-distribution detection. In *Advances in Neural Information Processing Systems*, pages 14707–14718, 2019.

[10] Lukas Ruff, Jacob R Kauffmann, Robert A Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G Dietterich, and Klaus-Robert Müller. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 109(5):756–795, 2021.

[11] Robin Schirrmeister, Yuxuan Zhou, Tonio Ball, and Dan Zhang. Understanding anomaly detection with deep invertible networks through hierarchies of distributions and features. *Advances in Neural Information Processing Systems*, 33:21038–21049, 2020.

[12] Joan Serrà, David Álvarez, Vicenç Gómez, Olga Slizovskaia, José F Núñez, and Jordi Luque. Input complexity and out-of-distribution detection with likelihood-based generative models. *arXiv preprint arXiv:1909.11480*, 2019.

[13] Claude E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948. doi:10.1002/j.1538-7305.1948.tb01338.x.

[14] Mingtian Zhang, Andi Zhang, and Steven McDonagh. On the out-of-distribution generalization of probabilistic image modelling. *Advances in Neural Information Processing Systems*, 34, 2021.

[15] Mingtian Zhang, Andi Zhang, Tim Z Xiao, Yitong Sun, and Steven McDonagh. Out-of-distribution detection with class ratio estimation. *arXiv preprint arXiv:2206.03955*, 2022.