Falsehoods that ML researchers believe about OOD detection

Andi Zhang
Computer Laboratory
University of Cambridge
az381@cam.ac.uk

Damon Wischik
Computer Laboratory
University of Cambridge
djw1005@cam.ac.uk

Abstract

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 approach has been found to fail, in deep learning settings. In this paper, we list some falsehoods that machine learning researchers believe about density-based OOD detection. Many recent works have proposed likelihood-ratio-based methods to 'fix' the problem. We propose a framework, the *OOD proxy framework*, to unify these methods, and we argue that likelihood ratio is a principled method for OOD detection and not a mere 'fix'. Finally, we discuss the relationship between domain discrimination and semantics.

1 Introduction

We might expect that a neural network should produce reliable outputs when it is presented with data similar to that used in training, and that its outputs might be prone to error when it is presented with substantially different data. It is clearly desirable for a neural network to be able to detect the latter case. This is called the out-of-distribution (OOD) detection problem [4].

A naive approach to OOD detection is as follows: First, train a density model $p(x)$ to approximate the true distribution from which the training dataset is assumed to be drawn. If $p(x)$ is small at some particular novel input x, it indicates that there is little training data in the region around x, and that the model should therefore be unconfident.

This naive approach leads to a paradoxical result, as elegantly shown by Nalisnick et al. [9]. They found that if they train a generative model to learn the density $p(x)$ on CIFAR10, and then evaluate this trained $p(x)$ on two test sets, one from CIFAR10 and one from SVHN, then the test CIFAR10 scores are *lower* than those for SVHN. They found this paradoxical result in several other examples, and it is easy to replicate.

In this paper, we will argue that this result is not in fact paradoxical: that in fact the naive approach to OOD detection is based on several falsehoods, falsehoods which are readily demonstrated using basic probability and statistics. These falsehoods are

- 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.

We will also argue that some successful approaches to OOD detection in the literature (starting with Bishop [1]) are based on the likelihood ratio between *two* datasets rather than the density for a single dataset, and that the substantive differences are to do with how this second dataset is constructed.

ML Safety Workshop, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

2 Falsehoods

Here is a simple example to illustrate the problems with using $p(x)$ for OOD detection. 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.

This phenomenon does not seem to be limited to toy examples. Nalisnick et al. [9] believe it can hold in real-world datasets, and explains their finding that $p(x)$ fitted to CIFAR-10 data is no good for detecting OOD datapoints from SVHN: 'Our conclusion is that SVHN simply "sits inside of" CIFAR10—roughly same mean, smaller variance—resulting in its higher likelihood.'

Outlier detection *v.* OOD detection. It is entirely reasonable to use $p(x)$ to test whether a datapoint is an outlier. This is indeed the cornerstone of frequentist statistics—we reject the null hypothesis when the test statistic shows that the observed data is unlikely. It's a reasonable basis for anomaly detection, to say 'if $p(x)$ is low then label x as an outlier." [13]

But OOD detection isn't the same thing as outlier detection. In the two examples above, we implicitly used the phrase 'OOD detection' to mean "drawn from a specified other distribution", and this other distribution happened to include non-outlier points.

In conclusion, $p(x)$ is not suitable for comparing two distributions. In section 3, we will discuss how to compare two distributions properly.

Lack of samples? Why was the result of Nalisnick et al. [9] 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 this intuition breaks down in high dimensions. A well-known result says that, with high probability, samples drawn a from high-dimensional Gaussian lie in a thin annulus — "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! [10]

In other words, "lack of samples" should not be confused with "low pdf".

3 Likelihood ratio

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 frequentist terminology, given an observation x, consider the null hypothesis H_0 that x was drawn from p_{in}, and the alternative hypothesis H_1 that x was drawn from p_{out}. By the Neyman-Pearson lemma [11], when fixing type-I error $P(\text{reject } H_0|H_0 \text{ is true})$, the test with the smallest type-II error $P(\text{accept } H_0|H_0 \text{ is false})$ is the likelihood ratio test. This implies that using likelihood ratio as a test score will optimise the area under the receiver operating characteristic (AUROC), which is a popular OOD detection baseline suggested by Hendrycks and Gimpel [4].

A similar result holds under the Bayesian perspective. Let $C \sim \text{Bin}(1,\alpha)$, and let $X \sim p_{\text{in}}$ if $C = 0$ and $X \sim p_{\text{out}}$ if $C = 1$. Given an observed value x,

$$\mathbb{P}(C=1|x) = \frac{p(x|C=1)\,\mathbb{P}(C=1)}{p(x|C=1)\,\mathbb{P}(C=1) + p(x|C=0)\,\mathbb{P}(C=0)} = \frac{1}{1+(1-\alpha)/(\alpha\,\mathit{LR})}$$

where $\mathit{LR} = p_{\text{out}}(x)/p_{\text{in}}(x)$ is the likelihood ratio. Since $\mathbb{P}(C=1|x)$ is an increasing function of *LR*, we might set a threshold θ and decide $C = 1$ if $\mathit{LR} > \theta$.

We have shown that the likelihood ratio is an optimal choice from both frequentist and Bayesian perspectives. However, it is hard to obtain p_{out}. In the next section, we introduce some practical works that propose proxies for p_{out}.

4 OOD proxies

In practice, we typically do not have an explicit p_{out} distribution. 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. This expresses the intuitive idea that p_{out} should spread out widely in a large area. Given x, the likelihood ratio between $p_{\text{in}}(x)$ and a constant is proportional to $p_{\text{in}}(x)$, which is identical to the criterion $p(x)$ used by Nalisnick et al. [9] if we ignore the scale of the threshold. 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[1] (e.g. 80 Million Tiny Images [3]) will increase the anomaly detection performance. Here, the auxiliary OOD datasets play a role of the $p_{\text{out}}^{\text{proxy}}$. Hendrycks et al. [5] did not use the likelihood ratio as the criterion for OOD detection, they proposed to fine-tune the generative model by the loss

$$\max\{0, C - \log p(x_{\text{in}}) + \log p(x_{\text{out}})\}$$

where C is a the number of the pixels of the image, x_{in} is the in-distribution data and x_{out} is the out-of-distribution data. Then they keep using the likelihood $p(x)$ to detect OOD. Following their OOD proxy, Schirrmeister et al. [15] 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. [19] suggested that the likelihood ratio could be estimated by a binary classifier.

Background statistics. Ren et al. [12] observed that "the background of images confounds the likelihood of the generative models", and propose a method for OOD detection based on eliminating the effect of background. Assume that background and semantic components are generated independently, i.e. $p(x) = p(x_S)\,p(x_B)$ where x_S stands for semantics and x_B stands for background. Suppose we know this factorization for the in-distribution data, as well as for proxy OOD data which has been obtained by perturbing the in-distribution data in such a way as to preserve the background and lose the semantics. They propose using the likelihood ratio $p_{\text{in}}(x_S)/p_{\text{out}}^{\text{proxy}}(x_S)$ for OOD detection.

In practice, it's hard to see how we can learn this factorization into semantics and background. They propose instead that $p_{\text{in}}(x_B) \approx p_{\text{out}}^{\text{proxy}}(x_B)$, since we perturbed the data so as to preserve the background. Then their likelihood ratio becomes

$$LR(x) = \frac{p_{\text{in}}(x_S)}{p_{\text{out}}^{\text{proxy}}(x_S)} \approx \frac{p_{\text{in}}(x_S)\,p_{\text{in}}(x_B)}{p_{\text{out}}^{\text{proxy}}(x_S)\,p_{\text{out}}^{\text{proxy}}(x_B)} = \frac{p_{\text{in}}(x)}{p_{\text{out}}^{\text{proxy}}(x)}$$

which is exactly our general-purpose likelihood ratio criterion.

Input complexity. Serrà et al. [17] observed that realistic in-distribution images typically have higher complexity, and that higher-complexity images typically have low $p(x)$, and suggest this is why using $p(x)$ is not good for OOD detection. They propose compensating for this by using a score $S(x) = \log_2 p_{\text{in}}(x) + L(x)$ where L is a measure of image complexity: the number of bits when x is compressed by a universal compressor. They point out that this is effectively a likelihood ratio test, using an OOD proxy distribution $p_{\text{out}}^{\text{proxy}}(x) \propto 2^{-L(x)}$. Similar to Ren et al., they interpret $p_{\text{out}}^{\text{proxy}}$ as describing the background without specific semantics.

[1] To have a fair comparison in the benchmark introduced by Hendrycks and Gimpel [4], the auxiliary OOD dataset does not have any intersection with the test OOD dataset.

Local features. Zhang et al. [18] 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.

Label-based. Suppose we're trying to detect OOD inputs to a classifier which we've trained on an dataset of (x,label) pairs. Hendrycks and Gimpel [4] suggested using the predicted labels for OOD detection: for example, if $y(x)$ is the vector of class probabilities predicted for input x, they suggest labelling x as OOD if the entropy $H\big(y(x)\big)$ is above a threshold. This idea has been taken on by others [6–8, 14, 16]. We can interpret the entropy-based detector as a likelihood ratio test, comparing p_{in} to $p_{\text{out}}^{\text{proxy}}$ defined by

$$p_{\text{out}}^{\text{proxy}}(x) \propto e^{H(y(x))}\, p_{\text{in}}(x).$$

It's interesting to speculate what this proxy distribution might look like; we are not aware of any work on this.

5 Discussion

Semantics *v.* domain distinction. The works we have discussed [12, 15, 17–19] 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" discussed in section 3. In effect, what we propose can be thought of as *defining* semantics in terms of domains: the semantics of p_{in} are those features that are absent in $p_{\text{out}}^{\text{proxy}}$.

One case where there is a somewhat clearer understanding of semantics is with labelled training data: the labels surely capture *some sort* of useful semantics. Label-based semantics can be linked to domain distinction, as shown by our label-based OOD proxy described above.

Likelihood-ratio is not a hack. Most of the works introduced in section 4 use 'failure' or some similar words to describe the phenomenon reported by Nalisnick et al. [9]. 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 section 3, 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.

Generalisation of OOD proxies. According to section 4, it is important to design a proper OOD proxy such that the model is able to distinguish the in-distribution test set $\mathcal{D}_{\text{in}}^{\text{test}}$ from many different OOD test sets $\mathcal{D}_{\text{out1}}^{\text{test}}, \mathcal{D}_{\text{out2}}^{\text{test}}, \ldots$ In other words, we want an OOD proxy that can distinguish the in-distribution domain from other domains. We call this the generalisation of OOD proxy. Hendrycks et al. [5] indicated that using real auxiliary data (e.g. 80 Million Tiny Images [3]) as the OOD proxy has a better performance than using the augmented in-distribution data. We believe this is because the real auxiliary data is more similar to the OOD data or has a large intersection with the domains of the OOD test datasets. Investigating the generalisation of OOD proxies is an open question, which we leave to future work.

References

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

[2] A. Blum, J. Hopcroft, and R. Kannan. *Foundations of data science*. Cambridge University Press, 2020.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. 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] D. Hendrycks and K. Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

[5] D. Hendrycks, M. Mazeika, and T. Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018.

[6] K. Lee, K. Lee, H. Lee, and J. Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems*, pages 7167–7177, 2018.

[7] S. Liang, Y. Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

[8] A. Malinin and M. Gales. Predictive uncertainty estimation via prior networks. In *Advances in Neural Information Processing Systems*, pages 7047–7058, 2018.

[9] E. Nalisnick, A. Matsukawa, Y. W. Teh, D. Gorur, and B. Lakshminarayanan. Do deep generative models know what they don't know? *International Conference on Learning Representations*, 2019.

[10] E. Nalisnick, A. Matsukawa, Y. W. Teh, and B. Lakshminarayanan. Detecting out-of-distribution inputs to deep generative models using a test for typicality. *arXiv preprint arXiv:1906.02994*, 5, 2019.

[11] J. Neyman and E. S. 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.

[12] J. Ren, P. J. Liu, E. Fertig, J. Snoek, R. Poplin, M. Depristo, J. Dillon, and B. Lakshminarayanan. Likelihood ratios for out-of-distribution detection. In *Advances in Neural Information Processing Systems*, pages 14707–14718, 2019.

[13] L. Ruff, J. R. Kauffmann, R. A. Vandermeulen, G. Montavon, W. Samek, M. Kloft, T. G. Dietterich, and K.-R. Müller. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 109(5):756–795, 2021.

[14] C. S. Sastry and S. Oore. Detecting out-of-distribution examples with gram matrices. In *International Conference on Machine Learning*, pages 8491–8501. PMLR, 2020.

[15] R. Schirrmeister, Y. Zhou, T. Ball, and D. Zhang. Understanding anomaly detection with deep invertible networks through hierarchies of distributions and features. *Advances in Neural Information Processing Systems*, 33:21038–21049, 2020.

[16] M. Sensoy, L. Kaplan, and M. Kandemir. Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing Systems*, pages 3179–3189, 2018.

[17] J. Serrà, D. Álvarez, V. Gómez, O. Slizovskaia, J. F. Núñez, and J. Luque. Input complexity and out-of-distribution detection with likelihood-based generative models. *arXiv preprint arXiv:1909.11480*, 2019.

[18] M. Zhang, A. Zhang, and S. McDonagh. On the out-of-distribution generalization of probabilistic image modelling. *Advances in Neural Information Processing Systems*, 34, 2021.

[19] M. Zhang, A. Zhang, T. Z. Xiao, Y. Sun, and S. McDonagh. Out-of-distribution detection with class ratio estimation. *arXiv preprint arXiv:2206.03955*, 2022.