What 'Out-of-distribution' Is and Is Not

Sebastian Farquhar
University of Oxford
sebfar@deepmind.com

Yarin Gal
OATML, Computer Science Department
University of Oxford

Abstract

Researchers want to generalize robustly to 'out-of-distribution' (OOD) data. Unfortunately, this term is used ambiguously causing confusion and creating risk—people might believe they have made progress on OOD data and not realize this progress only holds in limited cases. We critique a standard definition of OOD—difference-in-distribution—and then disambiguate four meaningful types of OOD data: transformed-distributions, related-distributions, complement-distributions, and synthetic-distributions. We describe how existing OOD datasets, evaluations, and techniques fit into this framework. We provide a template for researchers to carefully present the scope of distribution shift considered in their work.

1 Introduction

While the idea of adapting to distribution shift is old, perhaps surprisingly, researchers have only recently aimed to tackle 'out-of-distribution' (OOD) data following a call for concrete work on technical AI safety [Amodei et al., 2016].[1] Since then, over 10,000 papers are catalogued by Google Scholar on the topic which is central in work towards safer ML systems [Hendrycks et al., 2022b].

We want models that generalize or are robust to OOD data, or which can detect it. Sometimes, in order to achieve that, people train on OOD data. It is hard to define OOD data [D'Angelo and Henning, 2021], and on the rare occasions when it is defined it is usually with respect to a difference in distribution between the training data and some other data distribution [Lakshminarayanan, 2020, Tran et al., 2020]. Specifically, the standard definition is that, with respect to some reference data distribution $p_{\text{data}}(x, y)$[2] with $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, a target distribution $q(x, y)$ is OOD if and only if

$$p_{\text{data}}(x, y) \neq q(x, y). \tag{1}$$

This is a natural definition from the perspective of statistical learning theory [Vapnik, 1998], whose guarantees for model performance often assume that the target and reference distributions are equal. But, we argue, this definition is unhelpful in practice: it is too general and too demanding.

When people talk about OOD data they are talking about at least four different things. We distinguish four kinds of OOD data distributions which are special cases of the standard definition:

- **Transformed-distributions:** changed from the original by some set of transformations;
- **Related-distributions:** similar to the original in a way determined by use-context;
- **Complement-distributions:** (part of) the complement to the original distribution;
- **Synthetic-distributions:** distributions that are easy to sample but unrelated to the original.

These categories are orthogonal to categories like covariate shift and concept drift. We also note that in some cases it is useful to think of *data* as being OOD independently of a target distribution,

[1] See appendix A historical usage of the term 'OOD'.

[2] Our notation assumes supervised learning with input/output pairs, but all claims trivially cover, e.g., $p_{\text{data}}(x)$ and random variables with arbitrary dimensionality and event-space.

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

OOD category	Example Causes	Example Methods	Example Evaluation
Transformed	Sensor noise Adversaries	Data augmentation Adversarial training	ImageNet/CIFAR-C [Hendrycks and Dietterich, 2019] Adversarial Examples [Goodfellow et al., 2015]
Related	Deployment Task-shift Adversaries	Multi-task training Inductive priors	SVHN $\leftrightarrow$ MNIST [Nalisnick et al., 2019] CIFAR-10 $\rightarrow$ CIFAR-100 ('near-OOD') [Winkens et al., 2020] ImageNet-A and -O [Hendrycks et al., 2021] Species [Hendrycks et al., 2022a]
Complement	Open-set Adversaries	Forwards transfer Zero-shot learning Open set recognition	Split-MNIST/CIFAR [Kirkpatrick et al., 2017]
Synthetic	Rare	Negative sampling	Not typically used for evaluation.

Table 1: Transformed- and related-distributions are the most common naturally occurring kinds of OOD data. Transformations have well-specified functional form while related-distributions are possible for system designers to reason about. Adversaries are a common cause of many distribution shifts. Synthetic-distributions are very rare in practice, and are usually not evaluated on, but are fairly often used during training—e.g., by sampling uniform training data as a form of negative sampling.

but this is hard. Moreover, OOD data is emphatically *not* the same as outliers or anomalies—both of which can come from the reference distribution. OOD detection is not just a synonym for outlier or anomaly detection, despite papers often conflating these concepts (e.g., [Hendrycks et al., 2019, Winkens et al., 2020, Schirrmeister et al., 2020, Tran et al., 2022, Hendrycks et al., 2022a]).

Being imprecise about what OOD means creates problems for research and practice. One can accidentally motivate a problem with one kind of OOD data, solve it using another, and evaluate your method on a third! Being imprecise might make some kinds of OOD robustness seem possible that are theoretically intractable. For this reason, we propose a template in section 4 for describing OOD data use which should help avoid errors and make it easier to understand where work applies.

2 Failures of the Standard Definition of OOD

Statistical learning theory motivates understanding OOD as when a reference and target data distribution are not equal-in-distribution. Unfortunately, this is not a very helpful definition of OOD.

The definition is far too general. Suppose, for example, that the reference distribution is the probability distribution that generated the ImageNet dataset. Imagine you randomly duplicate just one of the datapoints in that dataset. The new dataset is drawn from a target distribution that is not equal-in-distribution to the reference dataset. This target distribution is therefore OOD with respect to the reference distribution *even though every single datapoint in the target is identical to a point in the reference distribution*. This is clearly not what we mean when we consider the problem of OOD detection, generalization, or robustness.

The definition is also far too demanding. Consider that there are infinitely many distributions that are not equal-in-distribution to any reference distribution (for one thing, consider the distributions whose event-spaces are different!). Assuming that we do not have extra information about which distributions are likely to be the target distribution, picking a robustness strategy is equivalent to picking a decision-rule for performing well on a distribution chosen uniformly-at-random. But this problem is well understood—it is impossible for any robustness strategy to out-perform any other in expectation by the no-free-lunch theorems [Wolpert and Macready, 1997]. Of course, we *can* do better than random when designing robustness strategies: this reveals that there is something in our intuitive conception of OOD which is missing from this definition.

3 Defining 'Out-of-Distribution'

In actual fact, people have a lot in mind when they think about OOD. They are interested in a set of distributions that are either conceptually related to $p_{\text{data}}(x, y)$ or are useful in some other way.

Transformed-distribution. Transformed distributions are like the reference distribution except that they are transformed according to a well-specified function mapping. For example, it might

represent a rotation, reflection, or translation in x; or perhaps a function producing a certain type of adversarial example.

In simple cases, robustness and generalization methods can tackle this kind of OOD data by explicitly or implicitly learning the underlying symmetry, or augmenting using data generated according to the transformation. In more complex cases, we will not have a list of the transformations that might produce OOD data. In those cases, we may need to resort to some sort of meta-learning procedure that uncovers transformations that are likely to occur in the real world.

Robustness evaluations like ImageNet-C [Hendrycks and Dietterich, 2019] evaluate the performance of models against this kind of OOD data: all the images in the test set are simple transformations of reference images. It is extremely important to recognize that being robust to one family of transformations offers no guarantees about performing well against an arbitrary alternative transformation (similarly to the no-free-lunch argument above). Robustness for one transformation *might* also help with some other transformation, but understanding this requires a careful analysis of the transformations that your system is likely to actually face.

Related-distribution. For transformed-distributions we can specify a transformation that maps our reference distribution onto a new one. But sometimes we know that there are some distributions we are likely to care about because as *system designers* we have an intuition about important data distributions. For example, you might imagine that when you build a classifier for different models of car someone might also accidentally try to classify models of trucks because they misunderstand the intended use of the software. They might even try to classify dog breeds out of pure curiosity. Similarly, the specific shift created by deploying your system is a very common related distribution.

These are best thought of as distributions, p_{related}, that are related to p_{data}, but not simple transformations of p_{data}. Rather, the *system designer* needs to reason, within the context in which the software is most likely to be deployed, about what kinds of data might be used as inputs and how the system ought to respond. Being robust to well-specified transformations often will not provide good results on related distributions, and there is no theoretical reason to think they will in general.

In principle, one could automate part of that special-knowledge reasoning and learn something about the typical distribution of p_{related} in our world for given contexts. That means the task becomes learning what the related distributions are (perhaps as a generative model or by learning D) and using this to inform the model learned using p_{data}. When researchers improve generalization on related-distributions, we hypothesise that they are doing something like this.

Complement-distribution. Sometimes we care about something like the complement of p_{data}: things that are not in the reference distribution. This might mean that $p_{\text{complement}}$ covers only classes or inputs which do not appear in p_{data}. For example, this is the structure of some challenges in open-set recognition [Scheirer et al., 2013] or continual learning [Kirkpatrick et al., 2017]. Formally speaking, two distributions that cover the same event-space cannot be complements, but often the probability density of one distribution over the high-density regions of the other is low enough that they are complements for practical purposes.

The ability to predict robustly on the complement of the observed data is impossible in general: if it is a new class or region of data-space then any decision-rule is compatible with the evidence in that region. The only thing that makes it tractable is implicit typical relationships between p_{data} and $p_{\text{complement}}$, which can be meta-learned. For this reason, papers that do open-set recognition often actually evaluate with related-distributions (e.g., [Tran et al., 2022]). This is fair enough, one cannot perform well on the entire complement, but places implicit limitations on the success of the method.

Synthetic-distribution: Sometimes, we are interested in distributions for reasons that have nothing to do with their relationship to the reference distribution. For example, they may be extremely easy to sample from, like the uniform distribution. This makes them much easier to assume access to during training (unlike related-distributions, which are far more useful but intrinsically hard to access).

There are reasons one might want to *use* synthetic distributions, but one should be very careful about confusing them and the guarantees they allow with distributions that are relevant to the actual task.

3.1 Data vs. Distribution

Although the conceptual formulation of 'out-of-distribution' refers to *distributions* people often talk about OOD *data*. What does it mean for target *data* to be OOD?

It could mean that the target data are not in the support of the reference distribution. For example, if the reference distribution is over the event-space of all sequences of English words, then most French sentences would be OOD datapoints. But usually we also think that points which are merely very unlikely under the reference distribution are OOD. Also, points that are impossible in the reference distribution but very similar to a point in the actual dataset are probably not OOD: consider a 1-bit `float32` perturbation of a training point that is natively `float16`.

"Points that are unlikely under the reference distribution" is also not a good definition of OOD data. As Le Lan and Dinh [2021] argue, density is only well-defined up to a parameterization of the space. Reparameterizing the input-space can change the relative order of high- or low-density points.

I do not currently believe that the important concepts behind OOD can be captured without reference to a second target distribution and the character of the differences between the reference and target distribution.

Outliers, anomalies, and OOD. The desire to define which points are OOD highlights a connection between OOD detection and the related problems of anomaly or outlier detection. Anomalies and outliers are specifically datapoints, not distributions. In fact, a crucial difference between OOD data and outliers/anomalies is that outliers/anomalies are *generated* by the reference distribution. That is, it is entirely possible for in-distribution data to be an outlier/anomaly. The common conflation of outliers, anomalies, and OOD data is therefore a mistake and should be avoided.

4 Why It Matters to be Precise About Definitions of OOD

This is not just a curiosity. The way we talk about concepts matters, both because it guides our research and because it guides how people use our research. Talking about OOD as some natural concept encourages the belief among researchers and practitioners that OOD generalization/robustness/detection is a solvable problem. But, in general, it is not! The idea that one can handle 'distribution shifts' in all generality is an illusion—for any decision rule you come up with there are data distributions which make it perform worse than chance. Indeed, in expectation over all possible data distributions it will never perform better than chance [Wolpert and Macready, 1997].

The solution is to be precise. People actually have some clear ideas about what kinds of distribution shifts they are likely to face. Here is what we propose: a distribution shift template for research into OOD robustness/generalization/detection. Papers with these goals should fill in the following fields:

- **Intended scope:** these are the kinds of distribution shift we aim to tackle. For example, 'Transformed distributions: common camera artefacts, viewing angles, pose changes, variable viewpoint distance.'
- **Intervention scope:** these are the kinds of distribution shift the training method, regularization techniques, and algorithmic innovations can be expected to deal with. For example 'Transformed distributions: rotations, reflections, translations, resizing, stretching, blur.'
- **OOD leakage:** are any kinds of OOD data assumed available during training? If so, are these from a common source with evaluation OOD data? Are genuinely 'OOD' data going to actually be available in deployment?
- **Evaluation scope:** these are the kinds of distribution shift which the evaluation techniques are actually designed to test. For example, 'Related distributions: 90 classes of 32×32 pixel images which are not in the training distribution.'

Similarly, papers that propose OOD evaluations should explicitly state their scope.

These templates help in two main ways. First, as a researcher (and reviewer) glancing at this quickly should reveal if there is a mismatch between the goals of the paper, its methods, and its evaluation. A mismatch is not necessarily a problem, it just means more work must be done. Second, as a practitioner glancing at this template should reveal if the method described in the paper actually corresponds to your intended use case.

Existential Risk Discussion

While out-of-distribution (OOD) generalization, robustness, and detection have been discussed in works related to reducing existential risks from AI (e.g., [Amodei et al., 2016, Hendrycks et al., 2022b]) the truth is that the vast majority of distribution shifts are not directly related to existential risks. When thinking about whether or not research into OOD generalization/robustness/detection is useful for reducing existential risks from AI it is therefore important to understand whether we should expect 'good' OOD properties to generalize from one kind of distribution shift to another.

As a first step, this requires rejecting the idea that OOD is a natural category. After all, if I believe my method is 'better on OOD robustness' then my belief stops me from interrogating whether it is specifically better on the sorts of distribution shifts that are likely to be associated with existential risks. That belief is the main target of this paper: we flatly reject the idea that OOD is a natural category.

It is our further, though not-yet-empirically-proven, belief that OOD properties do not generalize well from one kind of distribution shift to another. We intend to address this in future work. If this claim is true, then it follows that (in order to be useful for mitigating existential risk) OOD research should focus on identifying key distribution shifts associated with existential risk and targeting those *specifically* rather than hoping that OOD generalization/robustness/detection is a general strategy for mitigating existential risk. An example of this sort of shift is the specific shift that happens when a machine learning system becomes able to reason about its own training process [Cotra, 2022]

References

Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete Problems in AI Safety. *arXiv*, pages 1–29, 2016.

J Andrew Bagnell. Robust Supervised Learning. *AAAI*, 2005.

Yoshua Bengio. Deep Learning of Representations for Unsupervised and Transfer Learning. In *Proceedings of ICML Workshop on Unsupervised and Transfer Learning*, pages 17–36. JMLR Workshop and Conference Proceedings, June 2012.

Yoshua Bengio, Frédéric Bastien, Arnaud Bergeron, Nicolas Boulanger–Lewandowski, Thomas Breuel, Youssouf Chherawala, Moustapha Cisse, Myriam Côté, Dumitru Erhan, Jeremy Eustache, Xavier Glorot, Xavier Muller, Sylvain Pannetier Lebeuf, Razvan Pascanu, Salah Rifai, François Savard, and Guillaume Sicard. Deep Learners Benefit More from Out-of-Distribution Examples. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 164–172. JMLR Workshop and Conference Proceedings, June 2011.

Ajeya Cotra. Without specific countermeasures, the easiest path to transformative AI likely leads to AI takeover. https://www.lesswrong.com/posts/pRkFkzwKZ2zfa3R6H/without-specific-countermeasures-the-easiest-path-to, 2022.

Francesco D'Angelo and Christian Henning. Uncertainty-based out-of-distribution detection requires suitable function space priors. November 2021.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples, March 2015.

Dan Hendrycks and Thomas Dietterich. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. *ICLR*, March 2019.

Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep Anomaly Detection with Outlier Exposure. *arXiv:1812.04606 [cs, stat]*, January 2019.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural Adversarial Examples. *CVPR*, March 2021.

Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joe Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling Out-of-Distribution Detection for Real-World Settings. *International Conference on Machine Learning*, (arXiv:1911.11132), May 2022a.

Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. Unsolved Problems in ML Safety, April 2022b.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, March 2017.

Balaji Lakshminarayanan. Reliable Deep Anomaly Detection, 2020.

Charline Le Lan and Laurent Dinh. Perfect Density Models Cannot Guarantee Anomaly Detection. *Entropy*, 23(12):1690, December 2021. ISSN 1099-4300. doi: 10.3390/e23121690.

Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, and Balaji Lakshminarayanan. Detecting Out-of-Distribution Inputs to Deep Generative Models Using Typicality. *arXiv:1906.02994 [cs, stat]*, October 2019.

Walter J. Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E. Boult. Toward Open Set Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7): 1757–1772, July 2013. ISSN 1939-3539. doi: 10.1109/TPAMI.2012.256.

Robin Tibor Schirrmeister, Yuxuan Zhou, Tonio Ball, and Dan Zhang. Understanding Anomaly Detection with Deep Invertible Networks through Hierarchies of Distributions and Features. *Neural Information Processing Systems*, November 2020.

Dustin Tran, Balaji Lakshhminarayanan, and Jasper Snoek. Practical Uncertainty Estimation and Out-of-Distribution Robustness in Deep Learning, 2020.

Dustin Tran, Jeremiah Liu, Michael W. Dusenberry, Du Phan, Mark Collier, Jie Ren, Kehang Han, Zi Wang, Zelda Mariet, Huiyi Hu, Neil Band, Tim G. J. Rudner, Karan Singhal, Zachary Nado, Joost van Amersfoort, Andreas Kirsch, Rodolphe Jenatton, Nithum Thain, Honglin Yuan, Kelly Buchanan, Kevin Murphy, D. Sculley, Yarin Gal, Zoubin Ghahramani, Jasper Snoek, and Balaji Lakshminarayanan. Plex: Towards Reliability using Pretrained Large Model Extensions, July 2022.

Vladimir Vapnik. *Statistical Learning Theory*. Wiley, 1998.

Jim Winkens, Rudy Bunel, Abhijit Guha Roy, Robert Stanforth, Vivek Natarajan, Joseph R. Ledsam, Patricia MacWilliams, Pushmeet Kohli, Alan Karthikesalingam, Simon Kohl, Taylan Cemgil, S. M. Ali Eslami, and Olaf Ronneberger. Contrastive Training for Improved Out-of-Distribution Detection. *arXiv:2007.05566 [cs, stat]*, July 2020.

D.H. Wolpert and W.G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1(1):67–82, April 1997. ISSN 1089778X. doi: 10.1109/4235.585893.

A History of 'Out-of-distribution'

Researchers may be surprised to learn how recently the term 'out-of-distribution' has been introduced to describe the problem of generalization of neural networks to data that are not from the training distribution. For example, searching Google Scholar for the terms "out of distribution" and "neural network" before 2016 returns only 53 results. Of these, the vast majority are either: database errors which were actually published later; papers which merely cite a paper that uses the term in the title; papers which are using the phrase otherwise, e.g., "y is sampled out of distribution Y" or "purchases usually are made out of distribution warehouses". A manual check of all of these results reveals only four papers which discuss OOD data in any meaningfully relevant way, of which one is the journal version of another [Bagnell, 2005, Bengio et al., 2011, Bengio, 2012, Goodfellow et al., 2015].

Bagnell [2005] considers specifically the distribution shift that occurs when a system interacts with its environment (related-distribution). Bengio et al. [2011] and Bengio [2012] consider perturbed examples (transformed-distribution) and multi-task settings (related-distributions) and shows that

this sort of data can improve neural network learning. Goodfellow et al. [2015] asks why neural networks are so good at generalizing to points outside the training set given the presence of adversarial examples.

B Contrast to Standard Dataset Shift Definitions

Research into non-stationary distributions often distinguishes shift based on which part of $p_{\text{data}}(x, y)$ is changing. The main distinctions which are usually drawn are:

- Concept drift: $q(x, y) = p(x)q(y \mid x)$;
- Covariate shift: $q(x, y) = p(y \mid x)q(x)$;
- Prior shift: $q(x, y) = p(x \mid y)q(y)$

In concept drift, the input distribution is fixed, but the relationship between input and output changes in the target distribution. In covariate shift, the relationship between input and output is fixed, but the distribution of inputs changes. In prior shift, the relationship between output and input is fixed, but the distribution of the outputs changes.

These distinctions are unrelated to the ones we draw here. Any of these can be a transformed-, related-, complement-, or synthetic-distribution. For example, the transformation for a transformed distribution can easily be related to any of the conditional distributions or non-joint distributions that compose the three types of shift in this section. It can be useful to use both framings, possible at the same time, to characterize distribution shift.