
Dataset to Dataspace: A Topological-Framework to Improve Analysis of Machine Learning Model Performance

Henry Kvinge^{1,2} Colby Wight³ Sarah Akers³ Scott Howland³ Woongjo Choi³ Xiaolong Ma³ Luke Gosink³
Elizabeth Jurrus¹ Keerti Kappagantula³ Tegan H. Emerson^{1,4}

Abstract

As both machine learning models and the datasets on which they are evaluated have grown in size and complexity, the practice of using a few summary statistics to understand model performance has become increasingly problematic. This is particularly true in real-world scenarios where understanding model failure on certain subpopulations of the data is of critical importance. In this paper we propose a topological framework for evaluating machine learning models in which a dataset is treated as a “space” on which a model operates. This provides us with a principled way to organize information about model performance at both the global level (over the entire test set) and also the local level (on specific subpopulations). Finally, we describe a topological data structure, *presheaves*, which offer a convenient way to store and analyze model performance between different subpopulations.

1. Introduction

Advances in deep learning have resulted in major breakthroughs in a range of machine learning tasks (LeCun et al., 2015). The price of such advances have been dramatic increases in model size and complexity, fueled by ever larger training and test sets. As a consequence, it has become more challenging to understand model performance. In the literature, models are frequently judged based on a few statistics that are calculated over an entire test set. The use of such an evaluation scheme can lead to lack of robustness through underspecification (D’Amour et al., 2020). It can also obscure

serious model failures on certain subpopulations of the data (Buolamwini & Gebru, 2018; Yao et al., 2011; Recht et al., 2019).

In this paper we propose one approach to begin to address this issue, based on the observation that a “dataset” is rarely just a set of points. Instead, most datasets have a significant amount of metadata attached to them. In the supervised setting, this could include labels, but it might also include a wide variety of additional information. For example, from an image classification dataset we might have access to the location and time an image was taken, the kind of camera that was used or resolution of the image, or other objects or properties labeled in the image. It was observed in (Kvinge et al., 2021) that the presence of such metadata means that we should treat a dataset as a “space” rather than just a set, with proximity between datapoints based on similarities or differences in labels and metadata. In this paper we show how this paradigm can provide a framework for principled, fine-grained analysis of model performance.

Following (Kvinge et al., 2021), we begin by recalling how a topology, which encodes the bare essentials necessary to define a notion of space, can be built on top of a dataset. We remind the reader of *presheaves*, a data structure from topology that allows one to attach data locally (that is, to subpopulations of a dataset). We show how these constructions can be used to (i) systematically keep track of model performance across many subsets of related points and (ii) compare model performance on non-disjoint subsets. We finally show that *presheaves* allow us to make sense of a range of questions about a dataset which would otherwise be ill-defined. For example: “How does a model perform in the vicinity of a datapoint x ?” We conclude with an exploration of these ideas using a ResNet18 convolutional neural network (He et al., 2016) trained on the Caltech-UCSD Birds 200 dataset (Wah et al., 2011). We use the extensive image attribute metadata associated with this dataset to build a topology which reflects similarities in bird appearance that go beyond information contained in image labels alone. While many of the analytical approaches presented in this paper are actually rather simple and could have been developed without the underlying topological machinery, we

¹Pacific Northwest National Laboratory, Seattle, WA, USA

²Department of Mathematics, University of Washington, Seattle, WA, USA ³Pacific Northwest National Laboratory, Richland, WA, USA ⁴Department of Mathematics, Colorado State University, Fort Collins, CO, USA. Correspondence to: Henry Kvinge <henry.kvinge@pnl.gov>.

believe that building a precise mathematical framework in which to work makes it easier to conceive of and develop new approaches to evaluating model robustness across large and complex datasets.

2. Related work

Model robustness has become an increasingly important topic as deep learning models have begun to be deployed in a range of safety-critical applications. Some of this work has focused on the extent to which models are robust to perturbation or corruption of input (Hendrycks & Dietterich, 2019) or shifts in distribution (Hendrycks et al., 2020). The present work is inspired by a line of research investigating how models can lack robustness when they systematically fail on certain subpopulations of a dataset (Oakden-Rayner et al., 2020). There are a range of approaches to mitigating this phenomenon, including the use of novel loss functions or methods of training that promote more robust performance on specified or unspecified subpopulations (Duchi et al., 2020; Sohoni et al., 2020). Our work can be seen as a complementary approach to these methods, creating a framework in which one can systematically organize subpopulations of a dataset for the purposes of analyzing and mitigating undesirable model behavior.

In the last 10 years there has been a push to find ways of applying tools from the field of topology to questions in data science (Carlsson, 2009). Much of the resulting work, commonly known as *topological data analysis (TDA)*, has focused on developing methods of measuring the “shape” of point clouds via notions from topology such as homology (Edelsbrunner et al., 2000; Zomorodian & Carlsson, 2005). This is distinct from the present work which takes a more combinatorial approach to topology, choosing to build a finite topology (Barmak, 2011) induced by metadata.

Presheaves have only recently begun to be applied to problems in machine learning and data science. They have, for example, been used for uncertainty quantification in geolocation (Joslyn et al., 2020), air traffic control monitoring (Mansourbeigi, 2017), learning signals on graphs (Hansen & Ghrist, 2019), and data fusion (Robinson, 2017). This is the first time that this data structure has been applied to the problem analyzing model performance.

3. Background

3.1. Topologies and presheaves

This paper will utilize two foundational constructions from mathematics (1) the notion of a topology and (2) the notion of a presheaf. Due to length limitations, in this paper we confine ourselves to a concise definition of each, noting that the former is an entire discipline in mathematics and the

latter is a ubiquitous tool that appears across many fields of mathematics. We urge the interested reader to consult (Munkres, 2014; Hatcher, 2002) for further information about topology and (Vakil, 2017) Part 1, Chapter 2 and (Bredon, 1997) for further information about presheaves.

As a consequence of the way we will use topology in this paper and the fact that all objects we deal with are finite, we give a non-standard definition of a topology based on the concept of a subbasis.

Definition 3.1. Let X be a finite set and let $B = \{U_i\}_{i \in I}$ be a collection of subsets of X indexed by some set I . Assume that the union of all subsets in B is X . Let \mathcal{T} be the collection of all subsets of X that can be formed by some sequence of unions and intersections of elements of B (this includes the empty union). We call \mathcal{T} the *topology* induced by *subbasis* B .

The finite sets in \mathcal{T} are called *open sets* and are the finite analogue of open sets from more familiar topological spaces (e.g. \mathbb{R}). Note that it follows by construction that \mathcal{T} is (i) closed under unions, (ii) closed under intersections, and (iii) contains \emptyset (the empty union) and X (the union of all elements in B). These conditions happen to be the axiomatic definition of a topology of a finite set (Munkres, 2014). For $x \in X$, a *neighborhood of x* is any open set $U \in \mathcal{T}$ that contains x . We think of U as capturing some notion of the area “around x ”.

A presheaf is a structure that sits on top of a topological space and allows one to systematically (i) assign data to open sets and (ii) compare the data sitting on intersecting open sets.

Definition 3.2. Let X be a finite set with topology \mathcal{T} . Then a *presheaf* \mathcal{F} on \mathcal{T} is a function that to each open set $U \in \mathcal{T}$ associates a set $\mathcal{F}(U)$ (the *space of sections over U*), along with a *restriction map* $\text{res}_{U,V}^{\mathcal{F}} : \mathcal{F}(U) \rightarrow \mathcal{F}(V)$ for each open set V such that $V \subseteq U$, subject to the following conditions.

1. For any $U \in \mathcal{T}$, the trivial restriction map $\text{res}_{U,U}^{\mathcal{F}} : \mathcal{F}(U) \rightarrow \mathcal{F}(U)$ is the identity function from $\mathcal{F}(U)$ to $\mathcal{F}(U)$.
2. For any open sets $W, V, U \in \mathcal{T}$ with $W \subseteq V \subseteq U$, $\text{res}_{V,W}^{\mathcal{F}} \circ \text{res}_{U,V}^{\mathcal{F}} = \text{res}_{U,W}^{\mathcal{F}}$.

Note that through its restriction maps, a presheaf shadows the fact that a topological space \mathcal{T} can be completely defined through set inclusion maps $V \hookrightarrow U$ for each pair $V, U \in \mathcal{T}$ with $V \subseteq U$. Restriction maps provide a way of transferring data collected on a larger region of X , U , to a smaller region, V .

For open set U , each element of $\mathcal{F}(U)$ is called a *section*.

Following (Robinson, 2017), the choice of a section $a_U \in \mathcal{F}(U)$, for each $U \in \mathcal{T}$, $\{a_U\}_{U \in \mathcal{T}}$, is called an *assignment*.

3.2. Datasets as Topological Spaces

Suppose that we are handed a dataset D . As described in Section 1, we can often use metadata or labels to identify multiple subsets of related points from D . For example, if elements of D have labels from set L , then we can form the subsets $\{U_\ell\}_{\ell \in L}$, where U_ℓ contains all those elements $x \in D$ with label ℓ . Alternatively, if there are scalar values associated with elements of D , then for any $a \in \mathbb{R}$ we can form the set $U_{\geq a}$ (respectively, $U_{\leq a}$) which consists of all those values $x \in D$ such the scalar value associated with x is greater than (resp. less than) a .

Denote a choice of such subsets from D by $S = \{U\}_{i \in I}$ where I is some index set. As described in (Kvinge et al., 2021) one can form a topology \mathcal{T}_S on D by taking S as a subbasis. Then \mathcal{T}_S encodes a notion of space on D that is informed by labels and other metadata from D . We note that even though \mathcal{T}_S is a finite topology, it will potentially be very large and many, if not most, open sets may be hard to interpret, arising from combinations of unions and intersections of elements from S . We advocate putting limits on the number of intersections and unions that are actually calculated in practice. Furthermore, depending on the application one may be more interested in intersections than unions and vice versa. In the toy experiments described in Section 5 for example, we limit ourselves to intersections that include at most two of the subbasis elements.

4. Encoding Model Performance as a Presheaf

4.1. The accuracy presheaf

In this section we describe a presheaf designed to store a model’s accuracy on different open sets of the dataset topology \mathcal{T}_S outlined in Section 3.2. To this end we assume that dataset D is associated with a classification task with label space L . We note that this section is meant to function as a template for how a range of performance statistics might be encoded as presheaves. We could have chosen to use, for example, loss, precision, recall, or F1-score with only minor modification. We end the section by describing a range of ways we can use the accuracy presheaf to analyze model performance.

We create a presheaf \mathcal{A} on the topology \mathcal{T}_S on dataset D above by setting $\mathcal{A}(U) = [0, 1]$ (the closed interval from 0 to 1) for each $U \in \mathcal{T}_S$ with $U \neq \emptyset$ and $\mathcal{A}(\emptyset) = \{0\}$. For $U, V \in \mathcal{T}_S$ with $V \subseteq U$, we let $\text{res}_{U,V}^{\mathcal{A}}$ be the identity map if $V \neq \emptyset$ and let $\text{res}_{U,\emptyset}^{\mathcal{A}}$ be the zero map otherwise. We call \mathcal{A} the *accuracy presheaf*. Thus assignments from \mathcal{A} consist of numbers $0 \leq a_U \leq 1$ attached to each open set $U \in \mathcal{T}_S$. Let $f : X \rightarrow L$ be a model that has been trained to predict the

labels on data coming from the same or similar distribution as D . The *accuracy assignment* $\{a_U^f\}_{U \in \mathcal{T}_S}$ associated with f , is then defined such that a_U^f is the accuracy of f on subset U of dataset D . For example, if D is a dataset with images of dogs, then a particular open set U might consists of all dogs that have spots and are small. The value a_U is then just the model’s accuracy on this subset.

Beyond examining individual accuracies coming from accuracy assignment $\{a_U^f\}_{U \in \mathcal{T}_S}$ (for example, on which open set does f achieve its highest or lowest accuracy), one can also compare how accuracies change as we move from a superset to a subset. That is, if $V \subset U$ are both open sets in \mathcal{T}_S then we can look at the difference: $|\text{res}_{U,V}^{\mathcal{A}}(a_U) - a_V|$, which in this case is equal to $|a_U - a_V|$ unless $V = \emptyset$ (we leave the restriction map notation in view of this being a template for other statistics that might require non-identity restriction maps). For example, how does a model’s performance change when we move from images of spotted dogs to images of spotted dogs that are small? Inspired by (Robinson, 2017), (Kvinge et al., 2021) introduced the notion of the local inconsistency of an assignment in an effort to measure the extent to which an assignment changes across related open sets. We define a modified version of assignment inconsistency more appropriate for studying statistics related to machine learning models. Let \mathcal{A} be the accuracy presheaf on topological space \mathcal{T}_S for dataset D . For non-negative integer k and $U \in \mathcal{T}$, the *local k -bounded inconsistency at $U \in \mathcal{T}_S$* of an assignment $A = \{a_U\}_{U \in \mathcal{T}}$ is defined as

$$\text{Incon}_k(U, A) := \max_{\substack{V \subseteq U, V \in \mathcal{T} \\ |U \setminus V| \leq k}} |\text{res}_{U,V}^{\mathcal{A}}(a_U) - a_V|.$$

By including the bounding value k , we avoid the situation where U is a very large subset of D and V is a very small subset of D . Large changes in model performance are more likely in such situations and may reflect statistical irregularities rather than model failure.

Note that by using the notion of proximity induced by \mathcal{T}_S , for any element $x \in D$, we can ask how f performs in different neighborhoods of x . Formally, we define the *maximal performance of f in a neighborhood of x* as:

$$a_{\max,x} := \max_{U \in \mathcal{T}_S, x \in U} a_U.$$

We define the *minimal performance of f in a neighborhood of x* , $a_{\min,x}$ analogously. As we show in Section 5, these statistics can help illuminate the factors influencing the performance of a model on an individual test example.

5. Experiments

To explore some of the ideas described above, we apply our topological framework to the *Caltech-UCSD Birds* 200 dataset (Wah et al., 2011). This dataset has 11, 788 images

Open set	Accuracy
primary color: black \cap rhinoceros auklet	39.13
rhinoceros auklet	43.33
bill length: same as head \cap bill color: orange	46.03
bill color: orange \cap shape: duck-like	70.37
throat color: black \cap shape: duck-like	71.82
bill shape: spatulate \cap bill color: orange	72.72

Table 1. A list of the 3 neighborhoods of the image in Figure 1 on which a ResNet18 model f achieves its lowest accuracies (top) and the 3 neighborhoods on which f achieves its highest accuracies (bottom), all on the Caltech-UCSD Birds 200 dataset.

of birds belonging to 200 different species. We denote the training (respectively, test) set for this dataset by D_{train} (resp. D_{test}). Critically for our task, this dataset also comes with 312 binary attributes that, along with the bird classes themselves, can be used to generate a topology subbasis S . Specifically, S consists of subsets U_i where i is some binary attribute for a bird in an image or a bird class. For example, is the eye color of the bird orange? Note that instances from the same class need not all have the same attributes. Even if a bird has orange eyes, if its eyes are not visible in an image, then it does not get included in U_i . As mentioned in Section 3.2, for this toy example we limit ourselves to at most one intersection of elements from the subbasis S (so we only use elements of the form U_i and $U_i \cap U_j$ for $U_i, U_j \in S$) and do not use intersections with less than 20 elements, so in particular we only consider a small subset of the full topology, \mathcal{T}_{birds} . We train a ResNet18 convolutional neural network (He et al., 2016) f on D_{train} starting from the Torchvision weights (Russakovsky et al., 2015) pretrained on ImageNet (Marcel & Rodriguez, 2010).

While across the entire Birds test set f achieves an accuracy of 0.571 (i.e. 57.1%), our analysis reveals that f has vastly different performance on different subpopulations. Surprisingly, out of the 11,798 open sets we considered, f achieved 1.000 accuracy on 16 of them and 0.000 on 18 of them. The majority of open sets with either the highest or lowest accuracy are the intersection of a class and an attribute. We note that among those open sets that are either a single attribute or an intersection of attributes, the bright colors of red, blue, and green seem to be associated with higher accuracy. For example “underparts color: grey \cap throat color: blue” (accuracy 1.000) or “head pattern: capped \cap nape color: red” (accuracy 0.963).

We also analyze the local inconsistency of the set “has throat color: yellow”, finding it to be quite high, at 0.546 (with threshold k set to 20). This underscores the fact that the



Figure 1. The rhinoceros auklet image on whose neighborhood we analyze the performance of ResNet18 model f in Table 1.

performance of f can differ significantly when we restrict to open subsets of a subpopulation. For example, if U corresponds to “has throat color: yellow” and V corresponds to subset “has breast color: yellow \cap American goldfinch” then $\text{res}_{U,V}^{\mathcal{F}}(a_U) - a_V = -0.338$ while if V corresponds to “has belly pattern: solid \cap common yellowthroat” then $\text{res}_{U,V}^{\mathcal{F}}(a_U) - a_V = 0.546$. We found greater inconsistency on subsets associated with attributes rather than classes, for example the local inconsistency for the class “vermillion flycatcher” was only 0.080.

Finally, we demonstrate how the notion of the performance of a model in the neighborhood of a point can be used to better understand why f handles a given input the way that it does. We choose the image, x , of the rhinoceros auklet shown in Figure 1. We see from Table 1 that the two neighborhoods where the model performs worst consists of those images that are characterized by being a rhinoceros auklet which does not have bright plumage. On the other hand, the model does well on neighborhoods characterized by either the shape of the bird or the bill shape/color (note though that the bill color is not exclusively found in top performing neighborhoods as indicated by the third worst performing open set containing x).

For a more complete analysis, with more confident conclusions, a larger scale investigation including more of the topology \mathcal{T}_{birds} would be required. We hope that the brief examples that we provide in this section help make the abstract constructions in this paper more clear.

6. Conclusion

In this paper we propose a new paradigm for model evaluation which is guided by the idea that a model’s test set should be handled as a space rather than just a set. We show how this makes precise notions of “local” vs. “global” model performance, allowing a model trainer to better conceptualize the ways in which a model fails to be robust. We hope that this represents a first step toward bringing human’s natural spatial intuition to bear on the challenge of evaluating complex machine learning models.

References

- Barmak, J. A. *Algebraic topology of finite topological spaces and applications*, volume 2032. Springer, 2011.
- Bredon, G. E. Sheaves and presheaves. In *Sheaf Theory*, pp. 1–32. Springer, 1997.
- Buolamwini, J. and Gebru, T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pp. 77–91. PMLR, 2018.
- Carlsson, G. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.
- D’Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M. D., et al. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint arXiv:2011.03395*, 2020.
- Duchi, J., Hashimoto, T., and Namkoong, H. Distributionally robust losses for latent covariate mixtures. *arXiv preprint arXiv:2007.13982*, 2020.
- Edelsbrunner, H., Letscher, D., and Zomorodian, A. Topological persistence and simplification. In *Proceedings 41st annual symposium on foundations of computer science*, pp. 454–463. IEEE, 2000.
- Hansen, J. and Ghrist, R. Learning sheaf laplacians from smooth signals. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5446–5450. IEEE, 2019.
- Hatcher, A. *Algebraic Topology*. Cambridge University Press, 2002.
- He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.
- Hendrycks, D. and Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.
- Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. *arXiv preprint arXiv:2006.16241*, 2020.
- Joslyn, C. A., Charles, L., Depernoy, C., Gould, N., Nowak, K., Praggastis, B., Purvine, E., Robinson, M., Strules, J., and Whitney, P. A sheaf theoretical approach to uncertainty quantification of heterogeneous geolocation information. *Sensors*, 20:12:3418, 2020. <https://doi.org/10.3390/s20123418>.
- Kvinge, H., Jefferson, B., Joslyn, C., and Purvine, E. Sheaves as a framework for understanding and interpreting model fit. *ArXiv*, 2021.
- LeCun, Y., Bengio, Y., and Hinton, G. Deep learning. *nature*, 521(7553):436–444, 2015.
- Mansourbeigi, S. M. Sheaf theory approach to distributed applications: Analysing heterogeneous data in air traffic monitoring. *International Journal of Data Science and Analysis*, 3(5):34, 2017.
- Marcel, S. and Rodriguez, Y. Torchvision the machine-vision package of Torch. In *Proceedings of the 18th ACM international conference on Multimedia*, pp. 1485–1488, 2010. doi: 10.1145/1873951.1874254.
- Munkres, J. *Topology*. Pearson Education, 2014.
- Oakden-Rayner, L., Dunnmon, J., Carneiro, G., and Ré, C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. In *Proceedings of the ACM conference on health, inference, and learning*, pp. 151–159, 2020.
- Recht, B., Roelofs, R., Schmidt, L., and Shankar, V. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pp. 5389–5400. PMLR, 2019.
- Robinson, M. Sheaves are the canonical data structure for sensor integration. *Information Fusion*, 36:208–224, 2017. doi: 10.1016/j.inffus.2016.12.002.
- Russakovsky, O., Deng, J., Su, H., Krause, Jonathan and Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.
- Sohoni, N., Dunnmon, J., Angus, G., Gu, A., and Ré, C. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. *Advances in Neural Information Processing Systems*, 33, 2020.
- Vakil, R. The rising sea: Foundations of algebraic geometry notes, 2017. URL <http://math.stanford.edu/~vakil/216blog/FOAGnov1817public.pdf>.
- Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.
- Yao, B., Khosla, A., and Fei-Fei, L. Combining randomization and discrimination for fine-grained image categorization. In *CVPR 2011*, pp. 1577–1584. IEEE, 2011.

Zomorodian, A. and Carlsson, G. Computing persistent homology. *Discrete & Computational Geometry*, 33(2): 249–274, 2005.