The Semantic Shift Benchmark

Anonymous Authors[1]

Abstract

Most benchmarks for detecting semantic distribution shift do not consider how the semantics of the training set are defined. In other words, it is often unclear whether the 'unseen' images contain semantically different objects from the same distribution (e.g 'birds' for a model trained on 'cats' and 'dogs') or to a different distribution entirely (e.g Gaussian noise for a model trained on 'cats' and 'dogs'). In this work, we propose 'open-set' class splits for models trained on ImageNet-1K which come from ImageNet-21K. Critically, we structure the open-set classes based on semantic similarity to the closed-set using the WordNet hierarchy — we create 'Easy' and 'Hard' open-set splits to allow more principled analysis of the semantic shift phenomenon. Together with similar challenges based on FGVC datasets, these evaluations comprise the 'Semantic Shift Benchmark'.

1. Introduction

This report outlines the Semantic Shift Benchmark (SSB), a set of evaluations which isolate *semantic* shift, as compared to other forms of low-level distributional shifts. Specifically, given a training set, the SSB contains carefully curated class splits coming from unseen categories, where the unseen categories are structured for semantic similarity to the training set. In the case of ImageNet, we select categories from ImageNet-21K (Ridnik et al., 2021) and bin them based on their distance in the WordNet hierarchy to the regular ImageNet-1K classes.

The benchmark is motivated by the task of *open-set recognition* (OSR) which is the problem of detecting if a test-time instance comes from an unseen category (Scheirer et al., 2013; Neal et al., 2018; Bendale & Boult, 2016). However, we note that it is applicable to other problems which deal with semantic distribution shift (Han et al., 2019; Vinyals et al., 2016).

While this report focuses on a large-scale ImageNet benchmark, we also describe three evaluations on fine-grained datasets which have clear definitions of a semantic class. We describe these as they follow analogous principles to the ImageNet evaluation and may be of interest to the community, though we acknowledge they are not 'ImageNet-scale' evaluations. We benchmark strong maximum logit score (MLS) and ARPL (Chen et al., 2021) baselines on the new benchmark suite to motivate future research.

2. Benchmark Design

ImageNet. We introduce a large-scale evaluation for category shift, with open-set splits based on semantic distances to the training set. Specifically, we designate the original ImageNet-1K classes for the closed-set, and choose open-set classes from the *disjoint* set of ImageNet-21K-P (Ridnik et al., 2021). We exploit the hierarchical, tree-like semantic structure of the ImageNet database. For instance, the class 'elephant' can be labelled at multiple levels of semantic abstraction ('elephant', 'placental', 'mammal', 'vertebrate', 'animal'). Thus, for each pair of classes between ImageNet-1K and ImageNet-21K-P, we define the semantic distance between two classes as the total path distance between their nodes in the semantic tree. Specifically, we adopt the Leacock Chodorow Similarity (LCS) (Fellbaum, 1998), which measures lexical semantic similarity by finding the shortest path in the WordNet graph between two concepts, and scales that value by the maximum path length. The LCS between two concepts i and j is defined as:

$$s_{i,j} = -log(\frac{p(i,j)}{2d}), \tag{1}$$

where $p(i,j)$ denotes the shortest path length between i and j, and d denotes the taxonomy depth. We compute this distance for all i in the ImageNet-1K classes, and all j in the disjoint ImageNet-21K classes. Finally, we select 'Easy' and 'Hard' open-set splits by sorting the total distances to the closed-set and selecting two sets of 1000 categories. We provide examples of our open-set splits for the ImageNet dataset in Figure 1.

Fine-grained classification datasets. Consider the properties of fine-grained visual categorization (FGVC) datasets. These datasets are defined by an 'entry level' category, such as flowers (Nilsback & Zisserman, 2008) or birds (Wah et al., 2011). Within the dataset, all classes are variants of that single category, defining a single *axis of semantic variation*, *e.g.*, 'bird species' in the case of birds. Because the

Table 1. **OSR results on the Semantic Shift Benchmark.** We measure the classification accuracy and AUROC on the binary open-set decision. OSCR measures the trade-off between open and closed-set performance. OSR results are shown on 'Easy / Hard' splits.

Method	CUB			SCars			FGVC-Aircraft			ImageNet		
	Acc.	AUROC	OSCR	Acc.	AUROC	OSCR	Acc.	AUROC	OSCR	Acc.	AUROC	OSCR
ARPL+	85.9	83.5 / 75.5	76.0 / 69.6	96.9	**94.8 / 83.6**	**92.8 / 82.3**	91.5	87.0 / 77.7	83.3 / 74.9	78.2	**79.3 / 74.0**	66.3 / 63.0
MLS	**86.2**	**88.3 / 79.3**	**79.8 / 73.1**	**97.1**	94.0 / 82.2	92.2 / 81.1	**91.7**	**90.7 / 82.3**	**86.8 / 79.8**	**78.8**	78.7 / 72.8	**67.0 / 63.4**

axis of variation is well defined, it is reasonable to expect a classifier to learn it given a number of example classes — namely, to learn what bird species are and how they can be distinguished. This makes the analysis of semantic shift *better posed*.

We propose three FGVC datasets for the SSB: CUB (Wah et al., 2011), Stanford Cars (Krause et al., 2013), FGVC-Aircraft (Maji et al., 2013). These datasets come with labelled attributes (*e.g.*, `has_bill_shape::hooked` in CUB), which can be used to characterize the differences between classes and the degree of semantic shift, and hence the difficulty of the open-set problem.

Key differences from prior work Most existing semantic shift evaluations operate on small scale datasets (Neal et al., 2018; Han et al., 2019). Compared to other large-scale evaluations, our main contribution is in our aim to explicitly capture *semantic novelty*. For instance, (Sun et al., 2021) considers Places365 and SUN to be semantically different to ImageNet-1K. These evaluations do not consider how the 'OoD' images relate to the taxonomy or semantic classification system defined in the training data. As such, it is not clear whether the model is responding to a true semantic shift at test-time (which is the goal of OSR) or to some other low-level distribution shift. In contrast, our use of WordNet to identify easy and hard open-set examples respects the classification system used to build the ImageNet-1K dataset.

We note that other works have used ImageNet-21K as examples of 'unseen categories' with respect to ImageNet-1K (Bendale & Boult, 2016; Kumar et al., 2021). However, we structure for semantic distance based on the WordNet hierarchy, which we suggest provides a more principled basis for analyzing 'semantic shift'. ImageNet-O (Hendrycks et al., 2021) is another dataset which selects individual 'out-of-distribution' images from ImageNet-21K by selecting adversarial OoD examples with a trained ResNet-50. However, respecting the idea of novel *categories*, we choose entire categories of images rather than only the most adversarially challenging instances. As a result, our proposed 'unseen class' splits have 50k images each — as compared to 2000 in ImageNet-O — and facilitate a truly ImageNet-scale test set (see Table 2 for all split sizes). We also suggest that identification of semantic novelty as defined by a human taxonomy is a distinct, and possibly independent, challenge to detecting adversarially OoD examples of current models.

Table 2. **Statistics of the Semantic Shift Benchmark.** We show '#Classes(#Test Images)' for the known classes, and for the 'Easy', 'Medium' and 'Hard' open-set classes.

Dataset	Known	Easy	Medium	Hard
CUB	100 (2884)	32 (915)	34 (1004)	34 (991)
Stanford Cars	98 (3948)	76 (3170)	-	22 (923)
FGVC-Aircraft	50 (1668)	20 (667)	17 (565)	13 (433)
ImageNet	1000 (50000)	1000 (50000)	-	1000 (50000)

3. Evaluation Protocol

We provide benchmark results on the Semantic Shift Benchmark for the task of open-set recognition. In Table 1 we evaluate the 'Maximum Logit Score' (MLS) and ARPL+ (Chen et al., 2021), which are two state-of-the-art methods of open-set recognition. For the 'known/unknown' class decision, we report AUROC as is standard practise, as well as accuracy to allow potential gains in open-set performance to be contextualized in the closed-set accuracy of a model. We also report Open-Set Classification Rate (OSCR) (Dhamija et al., 2018) which measures the trade-off between accuracy and open-set detection rate as a threshold on the confidence of the predicted class is varied. We report results on 'Easy' and 'Hard' splits for all datasets, combining 'Medium' and 'Hard' examples into a single bin when applicable.

For the ImageNet benchmark, we use a ResNet50 pre-trained on the ImageNet-1K dataset and evaluate on the unseen classes. We finetune the ARPL model from a pre-trained ImageNet-1K checkpoint. In fine-grained classification, it is standard to pre-train models on ImageNet-1K. This is unsuitable for the proposed fine-grained OSR setting, as ImageNet-1K contains overlapping classes with the proposed datasets. Instead, we pre-train the network on Places (Zhou et al., 2017) using MoCoV2 self-supervised weights (Chen et al., 2020; Zhao et al., 2021).

We find that careful consideration of the semantics of the open-set classes leads to harder splits significantly reducing OSR performance. This is in contrast to 'openness' (Scheirer et al., 2013), the current measure used to assess the difficulty of an OSR problem, dependent only on the ratio of the number of closed to open-set classes. On ImageNet, we find the harder split leads to 5-6% worse AUROC for both methods. We also experimented with randomly subsampling first 1K and then 10K open-set classes, finding that introducing more classes during evaluation only reduced open-set performance by around 0.6% ($\approx 10\times$ less than our proposed splits).

References

Bendale, A. and Boult, T. E. Towards open set deep networks. In *CVPR*, 2016.

Chen, G., Peng, P., Wang, X., and Tian, Y. Adversarial reciprocal points learning for open set recognition. *IEEE TPAMI*, 2021.

Chen, X., Fan, H., Girshick, R., and He, K. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020.

Dhamija, A. R., Günther, M., and Boult, T. E. Reducing network agnostophobia. In *NeurIPS*, 2018.

Fellbaum, C. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.

Han, K., Vedaldi, A., and Zisserman, A. Learning to discover novel visual categories via deep transfer clustering. In *ICCV*, 2019.

Hendrycks, D., Zhao, K., Basart, S., Steinhardt, J., and Song, D. Natural adversarial examples. *CVPR*, 2021.

Krause, J., Stark, M., Deng, J., and Fei-Fei, L. 3d object representations for fine-grained categorization. In *International IEEE Workshop on 3D Representation and Recognition (3dRR)*, 2013.

Kumar, P., Anubhav, Shrivastava, A., and Kong, S. Open world vision challenge. *CVPR Workshop on Open World Vision*, 2021.

Maji, S., Rahtu, E., Kannala, J., Blaschko, M., and Vedaldi, A. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013.

Neal, L., Olson, M., Fern, X., Wong, W.-K., and Li, F. Open set learning with counterfactual images. In *ECCV*, 2018.

Nilsback, M.-E. and Zisserman, A. Automated flower classification over a large number of classes. In *Proceedings of the Indian Conference on Computer Vision, Graphics and Image Processing*, 2008.

Ridnik, T., Ben-Baruch, E., Noy, A., and Zelnik-Manor, L. Imagenet-21k pretraining for the masses. *ArXiv e-prints*, 2021.

Scheirer, W. J., Rocha, A., Sapkota, A., and Boult, T. E. Towards open set recognition. *IEEE TPAMI*, 2013.

Sun, Y., Guo, C., and Li, Y. React: Out-of-distribution detection with rectified activations. *NeurIPS*, 2021.

Vinyals, O., Blundell, C., Lillicrap, T., Kvukcuoglu, K., and Wierstra, D. Matching networks for one shot learning. In *NeurIPS*, 2016.

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.

Zhao, N., Wu, Z., Lau, R. W., and Lin, S. What makes instance discrimination good for transfer learning? In *ICLR*, 2021.

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., and Torralba, A. Places: A 10 million image database for scene recognition. In *IEEE TPAMI*, 2017.

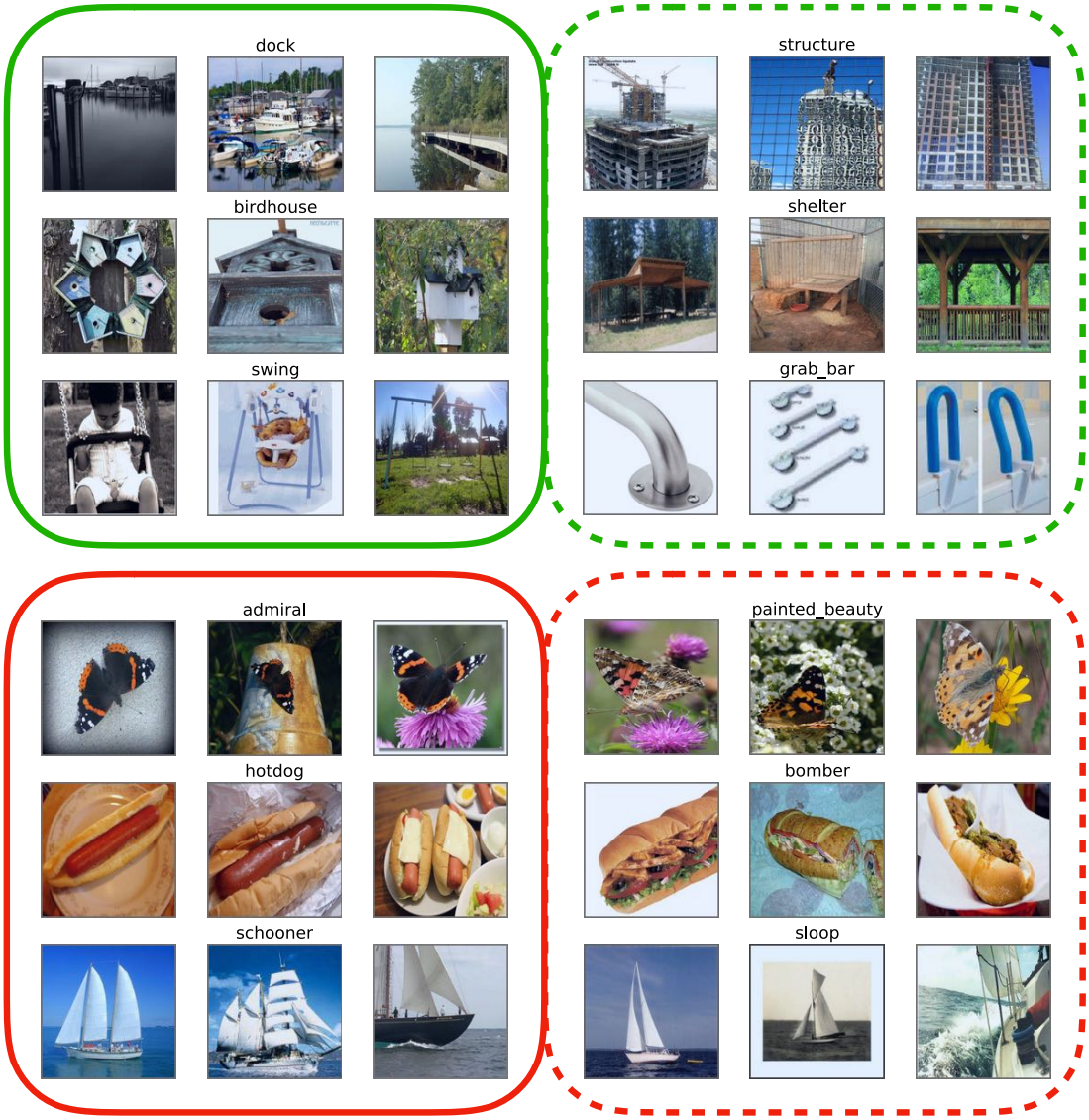

Figure 1. **Sample classes from closed and open-set splits for the ImageNet dataset.** We show 'Easy' (green/top) and 'Hard' (red/bottom) classes. Classes on the left (solid outline) are in the closed-set, while classes on the right (dashed outline) are in the open-set. In all cases, we show the *most similar* closed-set class to each open-set class.