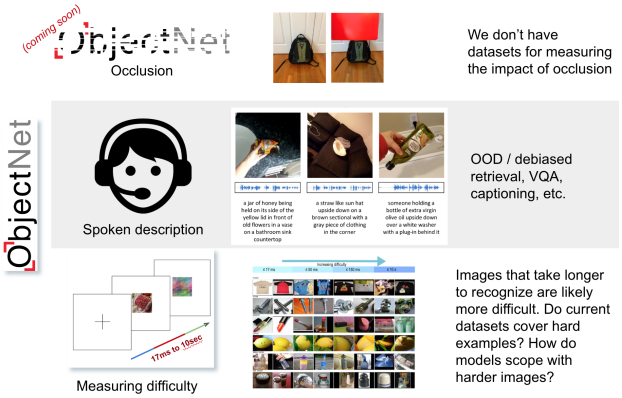
Growing ObjectNet: Adding speech, VQA, occlusion, and measuring dataset difficulty

David Mayo [1] **David Lu** [1] **Chris Zhang** [1] **Jesse Cummings** [1] **Xinyu Lin** [1] **Boris Katz** [1] **Jim Glass** [1] **Andrei Barbu** [1]

Abstract

Building more difficult datasets is largely an ad-hoc enterprise, generally relying on scale from the web or focusing on particular domains thought to be challenging. ObjectNet is an attempt to create a more difficult dataset systematically, by eliminating some biases that may artificially inflate machine performance. ObjectNet images are meant to decorrelate objects from their backgrounds, have randomized object orientations, and randomized viewpoints. ObjectNet appears to be much more difficult for machines. Spoken ObjectNet is a retrieval benchmark constructed from spoken descriptions of ObjectNet images. These descriptions are being used to create a captioning and VQA benchmark. In each case, large performance drops were seen. The next variant of ObjectNet will focus on real-world occlusions since it is suspected that models are brittle when shown partially-occluded objects. Using large-scale psychophysics on ObjectNet, we have constructed a new objective difficulty benchmark applicable to any dataset: the minimum presentation time for an image before the object contained within it can be reliably recognized by humans. This difficulty metric is well predicted by quantities computable from the activations of models, although not necessarily their ultimate performance. We hope that this suite of benchmarks will enable more robust models, provide better images for neuroscientific and behavioral experiments, and contribute to a systematic understanding of the dataset difficulty and progress in computer vision.



1. ObjectNet

ObjectNet (Barbu et al., 2019) is a dataset designed to challenge machine vision models by eliminating some of the confounds that make object recognition easier than it should be. Since current datasets are almost exclusively derived from images on the web, they have systematic biases. Images are selected to be pleasant for viewers. This leads to objects appearing on stereotyped backgrounds (plates are in kitchens rather than messy bedrooms) and in particular rotations (chairs are rarely shown on their sides). In addition, images are collected from common viewpoints (few images of plates straight on at eye-level appear anywhere). ObjectNet is designed to systematically eliminate these biases by explicitly collecting images with objects in particular backgrounds and rotations from particular viewpoints. We found a large performance gap between ImageNet and ObjectNet, even when considering only overlapping classes. Moreover, initially, it was thought that ObjectNet was not particularly more difficult for humans, as we describe below, after extensive psychophysics this is was found to be true; despite lacking the biases in ImageNet, ObjectNet is not significantly harder for humans. While recent progress in training multimodal models such as CLIP (Radford et al., 2021) and LiT (Zhai et al., 2022) has made significant advances in reducing the 50% performance drop between ImageNet and ObjectNet, much of the dataset remains to be solved, particularly the zero-shot portion requiring generalization to object classes which do not appear in ImageNet. We hope

[1]CSAIL & CBMM, MIT. Correspondence to: David Mayo <dmayo2@mit.edu>, Andrei Barbu <abarbu@mit.edu>.

Shift Happens workshop at the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

that the experience of ObjectNet encourages other groups to collect test-set-only datasets and forbidding training on their datasets in order to more robustly measure progress being made in computer vision.

We are contributing ObjectNet to the Shift Happens unified benchmark.

2. Spoken ObjectNet

The biases that ObjectNet controls for are prevalent in computer vision. To demonstrate this Spoken ObjectNet collected captions on Mechanical Turk for ObjectNet images (Palmer et al., 2021). These captions were used to create an image retrieval benchmark. Given an image, models must find the captions that best match that image. Spoken ObjectNet was collected from audio descriptions rather than written captions. Subjects are much more likely to speak at length than they are to write long captions. We found that the resulting captions were both longer and more complex. Spoken ObjectNet demonstrates a large performance drop, even larger than the one seen for object recognition.

3. Captioning and VQA ObjectNet

In upcoming work we are building new benchmarks for image captioning and VQA using Spoken ObjectNet. Biases seen in object recognition are often amplified in image captioning and VQA datasets. Certain situations are extremely rare, for example, unhappy beachgoers, so a caption like "happy people on the beach" is a very safe caption having recognized a beach and the presence of people, even if the model cannot recognize happy people. ObjectNet avoids these situations, hopefully leading to better datasets. To construct the captioning dataset, we are gathering five additional spoken captions per image, which will be used to score a generated caption. The VQA dataset is still being designed, with a focus on how to elicit a diverse set of questions.

4. ObjectNet Occlusion

The next version of ObjectNet will focus on occlusion. Few datasets focus on partially-occluded objects, and almost all of those add artificial occlusion. In preliminary experiments we have found that artificial occlusion is fairly easy for models to adapt to, while real-world occlusion is not. This is likely because artificial occlusion, and simple real-world occlusions, make detecting the occluder and segmenting it easy, while more complex real-world occlusions make determining the outline of the occluder hard to determine. We are in the process of designing a process by which to collect objects under complex occlusions, while controlling for both the amount of occlusion and the occlusion of critical object landmarks.

5. Objectively measuring dataset difficulty

In a publication in preparation, we report on a large-scale experiment collecting roughly 140,000 judgements from Mechanical Turk and in-lab subjects, related to the minimum viewing time of ObjectNet and ImageNet images. Subjects viewed cropped images at various presentation times, with masking, and were asked to perform a one out of 50 object classification task. We found that the distribution of image difficulties in ObjectNet and ImageNet was not particularly different. Moreover, it appears that several proxies computable from vision models (c-score (Jiang et al., 2021), prediction depth (Baldock et al., 2021), and adversarial robustness (Goodfellow et al., 2015)) predict the minimum viewing time with high accuracy. This has enabled us to compute the minimum viewing time distributions for many common datasets, finding that they too skew easy. Minimum viewing time provides an objective metric by which to measure datasets in computer vision, one that can be computed on the fly to calibrate dataset difficulty during the collection phase, and one that can be used to more rigorously evaluate progress made by the entire object recognition community.

Software and Data

ObjectNet is stored at `https://objectnet.dev` while Spoken ObjectNet is available at `https://github.com/iapalm/Spoken-ObjectNet`. Both datasets have a permissive license which allows for both commercial and academic use. As ObjectNet is a test set, training on ObjectNet is not allowed according on its license.

References

Baldock, R. J. N., Maennel, H., and Neyshabur, B. Deep Learning Through the Lens of Example Difficulty. *NeurIPS*, 2021.

Barbu, A., Mayo, D., Alverio, J., Luo, W., Wang, C., Gutfreund, D., Tenenbaum, J., and Katz, B. ObjectNet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *NeurIPS*, pp. 9448–9458, 2019. URL `objectnet.dev`.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and Harnessing Adversarial Examples. *ICLR*, 2015.

Jiang, Z., Zhang, C., Talwar, K., and Mozer, M. C. Characterizing Structural Regularities of Labeled Data in Overparameterized Models. *ICML*, 2021.

Palmer, I., Rouditchenko, A., Barbu, A., Katz, B., and Glass, J. Spoken ObjectNet: A bias-controlled spoken caption

dataset. *Interspeech*, 2021. URL https://arxiv.org/abs/2110.07575.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning Transferable Visual Models From Natural Language Supervision. *ICML*, 2021.

Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A., and Beyer, L. LiT: Zero-Shot Transfer with Locked-image text Tuning. *arXiv:2111.07991*, 2022.