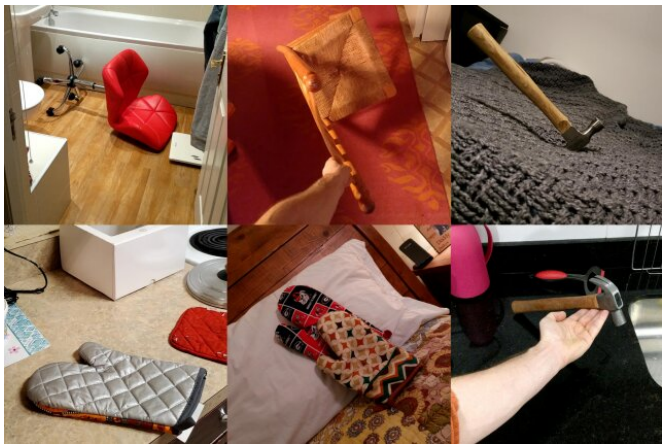
This object-recognition dataset stumped the world's best computer vision models

11 December 2019, by Kim Martineau



ObjectNet, a dataset of photos created by MIT and IBM researchers, shows objects from odd angles, in multiple orientations, and against varied backgrounds to better represent the complexity of 3D objects. The researchers hope the dataset will lead to new computer vision techniques that perform better in real life. Credit: Massachusetts Institute of Technology

Computer vision models have learned to identify objects in photos so accurately that some can outperform humans on some datasets. But when those same object detectors are turned loose in the real world, their performance noticeably drops, creating reliability concerns for self-driving cars and other safety-critical systems that use machine vision.

In an effort to close this performance gap, a team of MIT and IBM researchers set out to create a very different kind of object-recognition dataset. It's called ObjectNet, a play on ImageNet, the crowdsourced database of photos responsible for launching much of the modern boom in artificial intelligence.

Unlike ImageNet, which features photos taken from Flickr and other social media sites, ObjectNet features photos taken by paid freelancers. Objects are shown tipped on their side, shot at odd angles, and displayed in clutter-strewn rooms. When leading object-detection models were tested on ObjectNet, their accuracy rates fell from a high of 97 percent on ImageNet to just 50-55 percent.

"We created this dataset to tell people the object-recognition problem continues to be a hard problem," says Boris Katz, a research scientist at MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL) and Center for Brains, Minds and Machines (CBMM). "We need better, smarter algorithms." Katz and his colleagues will present ObjectNet and their results at the Conference on Neural Information Processing Systems (NeurIPS).

Deep learning, the technique driving much of the recent progress in AI, uses layers of artificial "neurons" to find patterns in vast amounts of raw data. It learns to pick out, say, the chair in a photo after training on hundreds to thousands of examples. But even datasets with millions of images can't show each object in all of its possible orientations and settings, creating problems when the models encounter these objects in real life.

ObjectNet is different from conventional image datasets in another important way: it contains no training images. Most datasets are divided into data for training the models and testing their performance. But the training set often shares subtle similarities with the test set, in effect giving the models a sneak peak at the test.

At first glance, ImageNet, at 14 million images, seems enormous. But when its training set is excluded, it's comparable in size to ObjectNet, at 50,000 photos.

"If we want to know how well algorithms will perform in the real world, we should test them on images that are unbiased and that they've never seen before," says study co-author Andrei Barbu, a

research scientist at CSAIL and CBMM.

A dataset that tries to capture the complexity of real-world objects

Few people would think to share the photos from ObjectNet with their friends, and that's the point. The researchers hired freelancers from Amazon Mechanical Turk to take photographs of hundreds of randomly posed household objects. Workers received photo assignments on an app, with animated instructions telling them how to orient the assigned object, what angle to shoot from, and whether to pose the object in the kitchen, bathroom, bedroom, or living room.

They wanted to eliminate three common biases: objects shown head-on, in iconic positions, and in highly correlated settings—for example, plates stacked in the kitchen.

It took three years to conceive of the dataset and design an app that would standardize the data-gathering process. "Discovering how to gather data in a way that controls for various biases was incredibly tricky," says study co-author David Mayo, a graduate student at MIT's Department of Electrical Engineering and Computer Science. "We also had to run experiments to make sure our instructions were clear and that the workers knew exactly what was being asked of them."

It took another year to gather the actual data, and in the end, half of all the photos freelancers submitted had to be discarded for failing to meet the researchers' specifications. In an attempt to be helpful, some workers added labels to their objects, staged them on white backgrounds, or otherwise tried to improve on the aesthetics of the photos they were assigned to shoot.

Many of the photos were taken outside of the United States, and thus, some objects may look unfamiliar. Ripe oranges are green, bananas come in different sizes, and clothing appears in a variety of shapes and textures.

Object Net vs. ImageNet: how leading object-recognition models compare

When the researchers tested state-of-the-art computer vision models on ObjectNet, they found a performance drop of 40-45 percentage points from ImageNet. The results show that object detectors still struggle to understand that objects are three-dimensional and can be rotated and moved into new contexts, the researchers say. "These notions are not built into the architecture of modern object detectors," says study co-author Dan Gutfreund, a researcher at IBM.

To show that ObjectNet is difficult precisely because of how objects are viewed and positioned, the researchers allowed the models to train on half of the ObjectNet data before testing them on the remaining half. Training and testing on the same dataset typically improves performance, but here the models improved only slightly, suggesting that object detectors have yet to fully comprehend how objects exist in the real world.

Computer vision models have progressively improved since 2012, when an object detector called AlexNet crushed the competition at the annual ImageNet contest. As datasets have gotten bigger, performance has also improved.

But designing bigger versions of ObjectNet, with its added viewing angles and orientations, won't necessarily lead to better results, the researchers warn. The goal of ObjectNet is to motivate researchers to come up with the next wave of revolutionary techniques, much as the initial launch of the ImageNet challenge did.

"People feed these detectors huge amounts of data, but there are diminishing returns," says Katz. "You can't view an object from every angle and in every context. Our hope is that this new dataset will result in robust computer vision without surprising failures in the real world."

More information: OnjectNet: objectnet.dev/

ImageNet: www.image-net.org/

teaching.

Provided by Massachusetts Institute of Technology