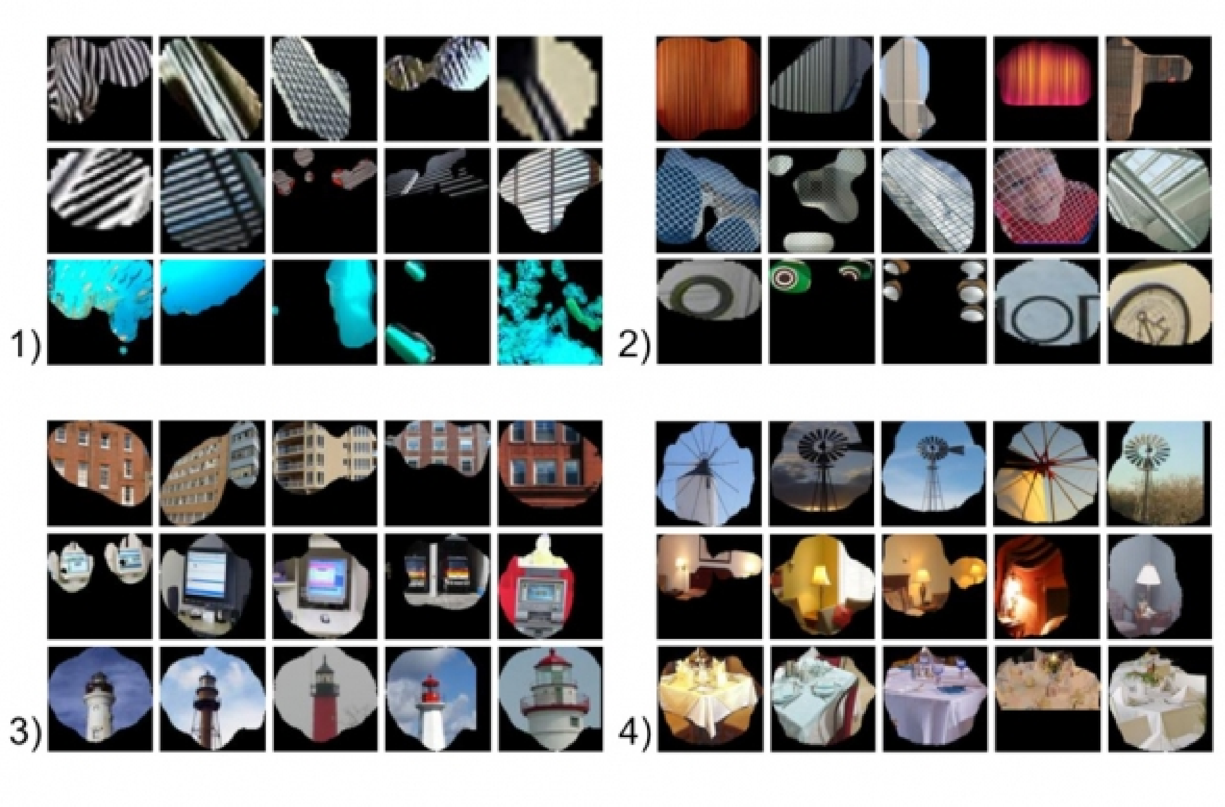
PHYS.ORG

System designed to label visual scenes according to type learns to detect specific objects

May 8 2015, by Larry Hardesty



The first layers (1 and 2) of a neural network trained to classify scenes seem to be tuned to geometric patterns of increasing complexity, but the higher layers (3 and 4) appear to be picking out particular classes of objects.

Object recognition—determining what objects are where in a digital image—is a central research topic in computer vision.

But a person looking at an image will spontaneously make a higher-level judgment about the scene as whole: It's a kitchen, or a campsite, or a conference room. Among computer science researchers, the problem known as "scene recognition" has received relatively little attention.

Last December, at the Annual Conference on Neural Information Processing Systems, MIT researchers announced the compilation of the world's largest database of images labeled according to scene type, with 7 million entries. By exploiting a machine-learning technique known as "deep learning"—which is a revival of the classic artificial-intelligence technique of neural networks—they used it to train the most successful scene-classifier yet, which was between 25 and 33 percent more accurate than its best predecessor.

At the International Conference on Learning Representations this weekend, the researchers will present a new paper demonstrating that, en route to learning how to recognize scenes, their system also learned how to recognize objects. The work implies that at the very least, scene-recognition and object-recognition systems could work in concert. But it also holds out the possibility that they could prove to be mutually reinforcing.

"Deep learning works very well, but it's very hard to understand why it works—what is the internal representation that the network is building," says Antonio Torralba, an associate professor of computer science and engineering at MIT and a senior author on the new paper. "It could be that the representations for scenes are parts of scenes that don't make any sense, like corners or pieces of objects. But it could be that it's objects: To know that something is a bedroom, you need to see the bed; to know that something is a conference room, you need to see a table and chairs. That's what we found, that the network is really finding these objects."

Torralba is joined on the new paper by first author Bolei Zhou, a graduate student in electrical engineering and computer science; Aude Oliva, a principal research scientist, and Agata Lapedriza, a visiting scientist, both at MIT's Computer Science and Artificial Intelligence Laboratory; and Aditya Khosla, another graduate student in Torralba's group.

Under the hood

Like all machine-learning systems, neural networks try to identify features of training data that correlate with annotations performed by human beings—transcriptions of voice recordings, for instance, or scene or object labels associated with images. But unlike the machine-learning systems that produced, say, the voice-recognition software common in today's cellphones, neural nets make no prior assumptions about what those features will look like.

That sounds like a recipe for disaster, as the system could end up churning away on irrelevant features in a vain hunt for correlations. But instead of deriving a sense of direction from human guidance, neural networks derive it from their structure. They're organized into layers: Banks of processing units—loosely modeled on neurons in the brain—in each layer perform random computations on the data they're fed. But they then feed their results to the next layer, and so on, until the outputs of the final layer are measured against the data annotations. As the network receives more data, it readjusts its internal settings to try to produce more accurate predictions.

After the MIT researchers' network had processed millions of input images, readjusting its internal settings all the while, it was about 50 percent accurate at labeling scenes—where human beings are only 80 percent accurate, since they can disagree about high-level scene labels. But the researchers didn't know how their network was doing what it was doing.

The units in a neural network, however, respond differentially to different inputs. If a unit is tuned to a particular visual feature, it won't respond at all if the feature is entirely absent from a particular input. If the feature is clearly present, it will respond forcefully.

The MIT researchers identified the 60 images that produced the strongest response in each unit of their network; then, to avoid biasing, they sent the collections of images to paid workers on Amazon's Mechanical Turk crowdsourcing site, who they asked to identify commonalities among the images.

Beyond category

"The first layer, more than half of the units are tuned to simple elements—lines, or simple colors," Torralba says. "As you move up in the network, you start finding more and more objects. And there are other things, like regions or surfaces, that could be things like grass or clothes. So they're still highly semantic, and you also see an increase."

According to the assessments by the Mechanical Turk workers, about half of the units at the top of the network are tuned to particular objects. "The other half, either they detect objects but don't do it very well, or we just don't know what they are doing," Torralba says. "They may be detecting pieces that we don't know how to name. Or it may be that the network hasn't fully converged, fully learned."

In ongoing work, the researchers are starting from scratch and retraining their network on the same data sets, to see if it consistently converges on the same objects, or whether it can randomly evolve in different directions that still produce good predictions. They're also exploring whether object detection and scene detection can feed back into each other, to improve the performance of both. "But we want to do that in a way that doesn't force the network to do something that it doesn't want to do," Torralba says.

"Our visual world is much richer than the number of words that we have to describe it," says Alexei Efros, an associate professor of computer science at the University of California at Berkeley. "One of the problems with object recognition and object detection—in my view, at least—is that you only recognize the things that you have words for. But there are a lot of things that are very much visual, but maybe there aren't easy describable words for them. Here, the most exciting thing for me would be that, by training on things that we do have labels for—kitchens, bathrooms, shops, whatever—we can still get at some of these visual elements and visual concepts that we wouldn't even be able to train for, because we can't name them."

"More globally," he adds, "it suggests that even if you have some very limited labels and very limited tasks, if you train a model that is a powerful model on them, it could also be doing less limited things. This kind of emergent behavior is really neat."

More information: "Object detectors emerge in deep scene CNNS." arxiv.org/pdf/1412.6856.pdf

"Learning Deep Features for Scene Recognition using Places Database." places.csail.mit.edu/

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT

research, innovation and teaching.

Provided by Massachusetts Institute of Technology

Citation: System designed to label visual scenes according to type learns to detect specific objects (2015, May 8) retrieved 28 April 2024 from https://phys.org/news/2015-05-visual-scenes-specific.html