Helping robots see the big picture

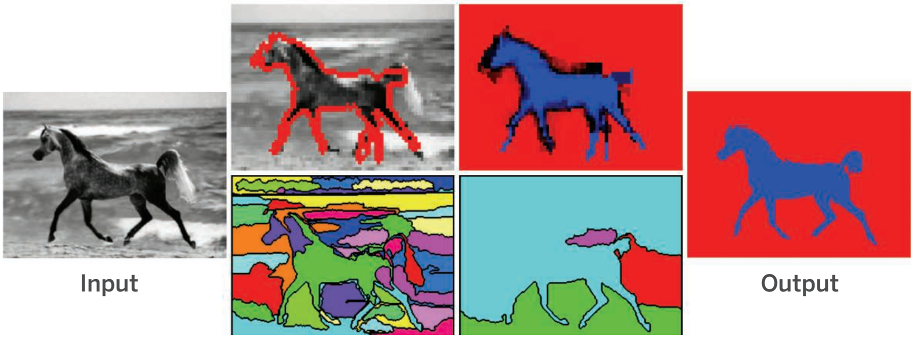
A computational approach called deep learning has transformed machine vision

By **John Bohannon**, *in San Francisco, California*

If you want to see the state of the art in machine vision, says artificial intelligence researcher D. Scott Phoenix, "you should watch the YouTube video of the robot making a sandwich." The robot in question is a boxy humanoid called PR2. It was built less than an hour away at Willow Garage in Menlo Park, California, one of the most influential robotics companies in the world. But Phoenix is being ironic. When PR2 finally manages to pick up a piece of bread, it drops the slice on the toaster; the bread caroms off, and a human rushes in to help. After stabbing a slice of salami with a fork, PR2 holds it in the air for what seems like an eternity. The sandwich does eventually get assembled, but it happens so slowly that the video is sped up 10-fold to make it watchable. And that was in an experimental kitchen in which "everything is carefully laid out," Phoenix says. "There's exactly one plate. Exactly one knife."

Looking beyond deep learning

A technique developed in the lab of Shimon Ullman uses a feedback system inspired by the brain's visual system to identify a horse in an image and locate it—missing only part of the tail.



Robots are clumsy because they struggle to make sense of all the data coming in from their cameras. "Vision is the biggest challenge," Phoenix says. Depending on their angle of view, objects can appear to have millions of different shapes. Change the lighting and each of those millions multiplies again—and this is the simplest case. A cluttered scene with overlapping objects is a nightmare. Although machines easily surpass human ability for certain constrained visual tasks, such as identifying a face among thousands of passport photos, as soon as they venture out "in the wild," as roboticists call the everyday human environment beyond the lab, machines flounder.

Two years ago, a powerful new computational technique, called deep learning, took the field of machine vision by storm. Inspired by how the brain processes visual information, a computer first learns the difference between everyday objects by creating a recipe of visual features for each. Those visual recipes are now incorporated into smart phone apps, stationary computers, and robots including PR2, giving them the capability to recognize what is in their environment. But roboticists worry that deep learning can't give machines the other visual abilities needed to make sense of the world—they need to understand the 3D nature of the objects, and learn new ones quickly on the fly—and researchers are already looking beyond deep learning for the next big advance.

PHOENIX CO-FOUNDED A STARTUP here called Vicarious, one of the myriad trying to capture human sight in code. Their optimism is surging. Over the past 2 years, deep learning has propelled machine vision by leaps and bounds. Where computers once struggled to detect the presence of something like a dog in a photo, with the help of deep learning they can now not only recognize a dog but even discern its breed.

"The theoretical side of deep learning was actually worked out decades ago," says Yann LeCun, a machine vision researcher at New York University in New York City who was one of its pioneers. He traces back its inspiration to the research of David Hubel and Torsten Wiesel, who shared a Nobel Prize in 1981 for research on biological vision. Working mostly with anesthetized cats in the 1960s and 1970s, Hubel and Wiesel discovered a hierarchical system of neurons in the brain that creates representations of images, starting with simple elements like edges and building up to complex features such as individual human faces. Computer scientists set about trying to capture the essence of this biological system. After all, LeCun says, "the brain was the only fully functioning vision system known."

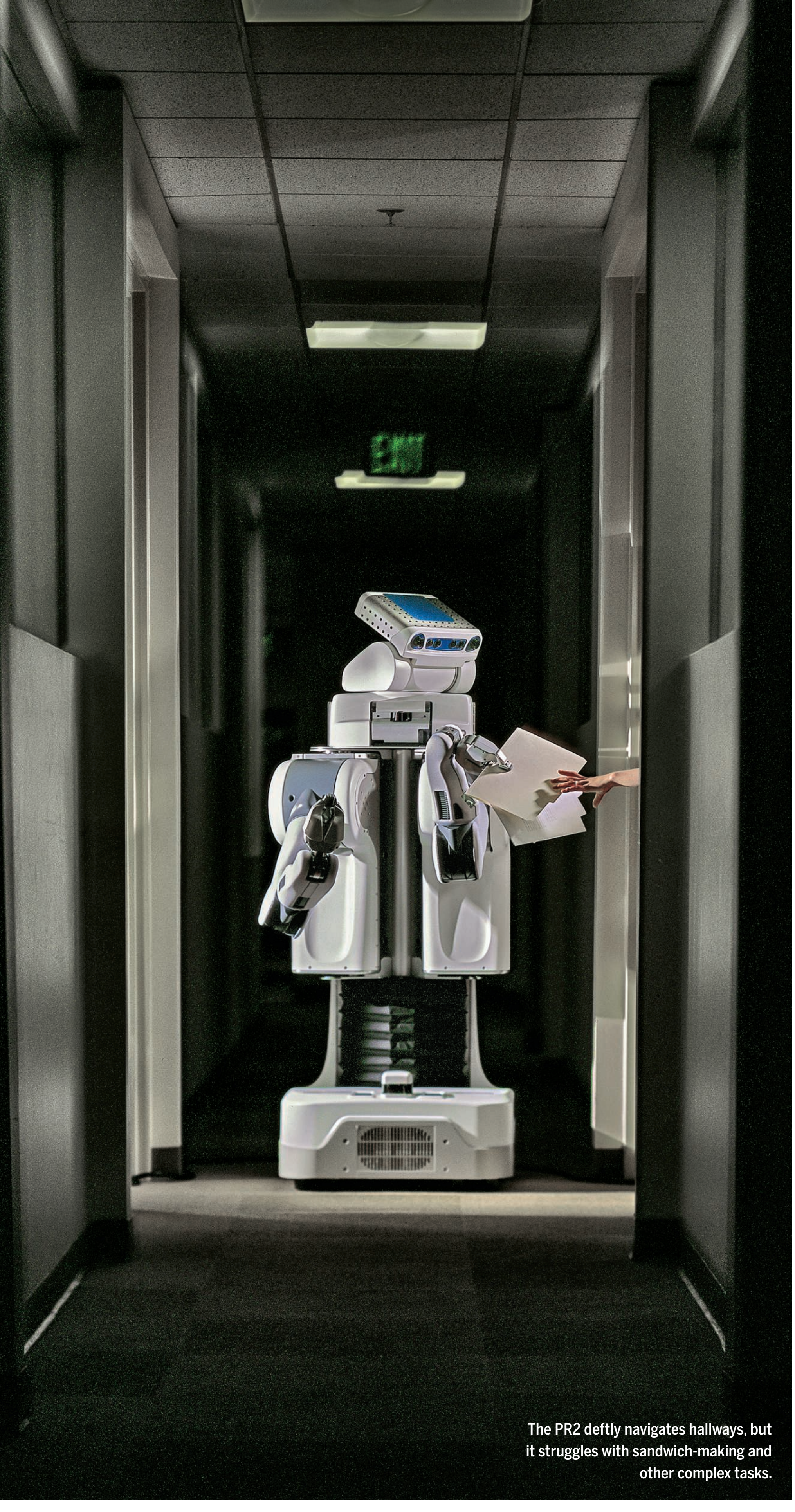
The deep learning architecture that emerged is called a deep convolutional neural network. Information flows between virtual neurons in a network. And like the real neurons in the brain's visual system, they are arranged in hierarchical layers that detect ever more complex features based on information from the previous layer. For example, the network would first break down a photo of a dog into edges between dark and light areas and then pass that information to the next layer for processing. By the time the last layer is reached, the system can apply a mathematical function to answer the question: Is this a dog or not a dog?

The problem was getting the dog-detecting function. "We just did not have the computers that we needed," LeCun says, nor the data. A network needed to process millions of labeled images of a dog to learn what a generic dog looks like, and in the 1980s even supercomputers did not have the speed or memory to handle this training. So researchers turned away from deep learning. Machine vision improved only incrementally—until 2012.

That year, a team led by computer scientist Geoffrey Hinton of the University of Toronto in Canada entered the ImageNet Challenge, an annual event in which competing computer programs must identify which objects —people, animals, vehicles—are present in thousands of photographs. Hinton's team used a deep convolutional neural network, trained on massive sets of labeled images. Unlike the computers of the 1980s, today's cheap computers have more than enough speed and memory for the calculations. Their system blew the competition out of the water.

"That changed everything," LeCun says. "The accuracy was so good that everyone in machine vision dropped what they were doing and switched to deep learning." Since then, billions of dollars have flooded into deep learning research. Hinton now develops deep learning applications at Google,

IMAGES: (LEFT) SHIMON ULLMAN; (RIGHT) MAX AGUILERA-HELLWEG



The PR2 deftly navigates hallways, but it struggles with sandwich-making and other complex tasks.

which is hoping to use the technique to create cars that can drive themselves. And at Facebook, where LeCun now heads the company's artificial intelligence efforts, a project called DeepFace aims to automatically identify any face in any photograph. (Whether the face's owner consents to being identified is another question.)

But amid the deep learning gold rush, some researchers are skeptical about the technique's prospects. Take PR2's epic struggle to make a sandwich. Deep learning enables the robot to recognize the objects around it—bread, salami, toaster—but it also needs to know exactly where those objects are in relation to its moving hand. Predicting where the bread will go when it drops requires physics. And what if the toaster is unplugged? The robot wouldn't have a clue why the bread wasn't toasting. "We have a long way to go before machines can see as well as humans," says Tomaso Poggio, who studies both machine and biological vision at the Massachusetts Institute of Technology in Cambridge.

The U.S. National Science Foundation (NSF) agrees. Poggio now heads an NSF-funded initiative called the Center for Brains, Minds and Machines. Most of the center's research is focused on understanding how human vision works and emulating it with computers. For example, Poggio says, "I can show a child a couple of examples of something and he will identify it again easily without having to train on millions of images." He calls this trick object invariance—a representation of an object that allows humans to identify it in any setting, from any angle, in any lighting—and his research focuses on capturing it as a computer algorithm.

Tech companies aren't waiting on the sidelines. Some are exploring new biologically inspired computer hardware (see p. 182). Phoenix and his colleagues at Vicarious are focusing on a software solution to visual intelligence. Last year, they announced that their algorithms had surpassed deep learning by cracking CAPTCHA, the visual puzzles made up of distorted letters that are used on websites to confound Web-crawling software. Vicarious has kept the details under tight wraps, but according to the company's co-founder, Dileep George, it is not based on deep learning at all. "We are working from how the human brain processes visual information."

When asked for something more tangible, George, like all entrepreneurs working furiously in secret, demurs. "It will be a few years before we pull it together," he says. And what is the goal? "A robot with the visual abilities of a 3-year-old." That would give robots the ability to do far more than make a sandwich. ■

Published by AAAS