Science | DOI:10.1145/2843532 Neil Savage

Seeing More Clearly

Computer understanding of images has improved rapidly, but true visual intelligence is still a long way off.

The photo shows a baby, dressed in a striped shirt and denim overalls, staring intently at the toothbrush he is grasping in his left hand while he pokes at it with his right. The caption underneath reads, "A young boy is holding a baseball bat."

The computer at Stanford University in California that generated that description failed in this case, but in others its captions were much more accurate, if not terribly exciting. The phrases "black and white dog jumps over bar," "little girl is eating piece of cake," and "baseball player is throwing ball in game" are all correct characterizations of the photos in question. Even if it occasionally mistakes a toothbrush for a baseball bat, or a ferret for a cat, the computer today can do a much better job of explaining what is happening in a photograph than it could even a couple of years ago. The field of computer understanding of images has made remarkable progress in the last few years.

"Things are moving really fast right now, because we are dealing with the earliest successes of object recognition and tasks," says Fei-Fei Li, director of the Vision Lab and the Artificial Intelligence Lab at Stanford, who developed the program that created those captions. Part of what has made such successes possible is the development of large sets of training data. Li, for instance, began in 2007 to build ImageNet, a collection that has grown to nearly a billion images hand-labeled by almost 50,000 people through Amazon Mechanical Turk, which provides small payments to people willing to perform such tasks. Microsoft, meanwhile, developed the Common Objects in Context (COCO) dataset, which contains more than 300,000 images, each labeled with five captions.

IMAGES COURTESY OF STANFORD.EDU

The other advance is that computers have become powerful enough to apply convolutional neural networks to the task of understanding images. Each neuron in the network is a filter that looks at a small image segment a few pixels wide and computes a value expressing how confident the computer is that a given object is within that segment. The segments overlap to cover the entire image, and the network repeats this process through many layers, with each layer's output providing input for the next layer. "With careful training, these things actually work very well," says Rob Fergus, a scientist in Facebook's AI Research Group in Menlo Park, CA. "Prior to these models, these recognitions systems didn't really work in practice."

Five years ago, Fergus says, recognition systems had an error rate of approximately 26%; today, that is down to 5%–6%. "It's not to say they can do visual recognition as well as a human, but in this slightly artificial setting, they do pretty well," he says. Eventually, computers should surpass humans. Thanks to their vast access to information, they should be able to always identify a breed of dog or the make and model year of a car, something most humans cannot manage.

Facebook—whose users upload 400 million images each day, according to Fergus—is very interested in automating image understanding. That social network wants to know whether to put a given photo into a user's newsfeed, based on what it shows or whether it depicts friends of the user; it also wants to detect and automatically delete objectionable content before people see it. One area Facebook is exploring is how to identify people whose faces are not visible, or partly obscured. While a human can look at a picture of President Obama with his face turned away from the camera, or Mahatma Gandhi with his head bowed in prayer, and instantly recognize them, a computer generally cannot.

Facebook researchers have developed a program called Pose Invariant Person Recognition (PIPER) that looks for "poselets"—a hand next to a hip, how a head and shoulders look from the back—that will tell it that it is seeing a person, even when it cannot identify a face. Given a starting image of an individual, the system was able to identify the same person in other photos 83% of the time. When a frontal image of a face was available, it improved the accuracy of Facebook's facial recognition software to 89%–94%.

"We aren't really at the point of [a computer] understanding what's going on in that image. At a glance, a person can figure that out."

Identifying discrete objects, however, is only a small part of understanding a scene. The next step toward visual intelligence is recognizing the relationship between those objects and noticing if action is taking place, and that is where attempts to write captions come in. Researchers at Microsoft trained their system by presenting the computer with images that were accompanied by human-generated captions. One advantage of that approach is that the humans who write the simple descriptions will tend to focus on the most important details. If the caption contains the word "horse," for instance, it is more likely that a horse will feature prominently in the picture, rather than be a small part of the background. The computer might also find a high correlation between the use of the word "riding" and images of a person on top of a horse.

Once trained, the computer follows a three-step process, says Xiaodong He, a researcher in Microsoft's Deep Learning Technology Center in Redmond, WA. First, it identifies objects within an image to create a list of words most likely to apply to the image. Then it uses those words to construct up to 500 candidate sentences. Finally, it ranks the sentences in order of likelihood, coming up with one that is most likely to describe what is in the picture. To do that, it creates a map of probabilities, called semantic vectors, ranking various words and phrases to decide which are more likely to go together. It also creates a similar ranking for words it has assigned to portions of the image. It compares the semantic values of the text to those of the image, and declares that those from each group that are closest to each other are most likely correct. "The overall picture should have the same semantic value as the description," says He.

In the MS COCO Image Captioning Challenge 2015, in which 15 groups from industry and academia competed to see who could do the best job of getting computers to generate descriptions, the Microsoft team ranked first according to a Turing-style test, in which judges were asked to determine whether the captions had been created by a human or a machine. Slightly more than 32% of the machine-generated captions were thought by the judges to have been written by people; only 68% of human-generated captions were attributed to humans by the judges, so the software is almost halfway toward passing the Turing test, He says. "If you looked at this problem two years ago, the outcome was almost garbage. It was so simple to tell which captions came from a human and which came from a computer," says He.

In another category based on the percentage of captions thought to be equal to or better than captions written by people, researchers from Google came out ahead; combining the scores left the two companies finishing in a tie.

Microsoft researchers taking a slightly different tack tied for third place with researchers from the University of Toronto, Canada, whose software also analyzes images and sentences to find the best match. The UT approach, developed with researchers from the University of Montreal, Canada, includes the concept of attention. The computer identifies the most important object in a given region of an image, then moves sequentially through the regions, with its decisions about what is most important in each influenced by what it saw in the others. Like Microsoft's system, it analyzes both the image and the sentence and tries to put together the two that best match.

While the program can correctly identify, say, that there are people in a boat in a particular photo, that is about as far as it goes, says Richard Zemel, a computer scientist at the University of Toronto. "What are they

actually doing? Are they rowing the boat? Are they falling out of the boat?" he says. "We aren't really at the point of understanding what's going on in that image. At a glance, a person can figure that out."

To move closer to such understanding, Zemel is working on training the computer to answer arbitrary questions about an image, such as "what color is the shirt?" or "what's in front of the sofa?" Answering such questions requires a more detailed description of an image than simply what is in it and where. "If you really understood the image, you could answer a question about it," Zemel says. "I think that's more indicative of true understanding." The work is still in its early stages, in part because the existing database of human-generated questions and answers about images the computer can learn from is not very large.

Li says true visual intelligence could be important in a lot of areas. Self-driving cars, for instance, have to be able to do more than simply drive down a road in a particular direction. If a computer-driven car comes across a construction zone, for instance, the new information it needs to proceed safely will be mostly visual, from such sources as officers directing traffic. "You have to read their gestures. You have to read detour signs. You have to see the orange cones," she says.

Electronic eyes with visual intelligence in a hospital could alert nurses to problems with a patient that the nurses missed, or remind them to wash their hands before they touch an IV. In an airport, they might identify an unattended backpack as a security issue.

Though recent progress has been rapid, Li believes it could slow as cientists try to tackle more complicated challenges than object recognition and move into reasoning about relationships among objects and actions. That will require building datasets that reflect a much more complex level of interconnectedness in the world. She compares it to the challenge of basing computer searches on natural language. Google does an excellent job when given discrete search terms, but ask it a long question, such as "give me the names of painkillers that do not have stomach side effects," and it stumbles.

Right now, computers are roughly as good at describing the content of images as a three-year-old human, Li says. "The complete level, on par with an adult, college-degreed human, I think is going to be a long way off." C

Further Reading

Fang, H., Gupta, S., Iandola, F., Srivastava, R.K., Deng, L., Dollár, P., Gao, J., He, X., Mitchell, M., Plattz J.C., Zitnick, C.L., Zweig, G.
From Captions to Visual Concepts and Back, ***Computer Vision and Pattern Recognition*****, Boston, MA, 2015.**

Zhang, N., Paluri, M., Taigman, Y., Fergus, R., and Bourdev, L.
Beyond Frontal Faces: Improving Person Recognition Using Multiple Cues, ***Computer Vision and Pattern Recognition*****, Boston, MA, 2015.**

Ren, M, Kiros, R., and Zemel, R.
Exploring Models and Data for Image Question Answering, ***Deep Learning Workshop at the International Conference on Machine Learning*****, Lille, France, 2015.**

Vinyals, O., Toshev, A., Bengio, S., and Erhan D.
Show and Tell: A Neural Image Caption Generator, ***Computer Vision and Pattern Recognition*****, Boston, MA, 2015.**

How We Teach Computers to Understand Pictures, Fei-Fei Li TED Talk. https://www.youtube.com/watch?v=40riCqvRoMs

Neil Savage is a science and technology writer based in Lowell, MA.

Milestones

American Academy Adds CS Fellows

Among the 181 Fellows and 16 Foreign Honorary Members recently elected to the American Academy of Arts and Sciences were seven computer scientists and a mathematician who is also a computer scientist.

Honorees in the Computer Sciences category were:

- **Sanjeev Arora, Charles C. Fitzmorris Professor of Computer Science at Princeton University.**
- **Susan T. Dumais, a Distinguished Scientist at Microsoft and affiliate professor at the University of Washington Information School.**
- **Laura M. Haas, IBM Fellow and director of the IBM Accelerated Discovery Lab at the IBM Almaden Research Center.**
- **Joseph Y. Halpern, a professor of computer science at Cornell University and administrator for the Computing Research Repository of arXIV.org.**
- **Maurice P. Herlihy, a computer scientist at Brown University.**
- **Ravindran Kannan, Principal Researcher at Microsoft Research Labs, Bangalore, India, and first adjunct faculty of the Computer Science and Automation Department of the India Institute of Science.**
- **Nicholas W. McKeown, a professor in the electrical engineering and computer science departments of Stanford University.**

In addition, 2007 ACM A.M. Turing Award laureate Joseph Sifakis, a computer scientist who works for CNRS at the VERIMAG laboratory, and as coordinator of Artist2, the European Network of Excellence for research on embedded Systems, was named Foreign Honorary Member in the Computer Sciences category.

New Fellows in the Mathematics, Applied Mathematics, and Statistics category included László Babai, a professor in the departments of Computer Science and Mathematics of the University of Chicago.

GUGGENHEIM FOUNDATION NAMES COMPUTER SCIENTISTS AS FELLOWS

The John Simon Guggenheim Memorial Foundation recently awarded Fellowships to 175 scholars, artists, and scientists in the U.S. and Canada on the basis of prior achievement and exceptional promise. Among these were two computer scientists, Vincent Conitzer and Krishna V. Palem.

The research of Conitzer, who is Sally Dalton Robinson Professor of Computer Science and professor of economics at Duke University, focuses on computational aspects of microeconomics, in particular game theory, mechanism design, voting/social choice, and auctions, using techniques from, and including applications to, artificial intelligence and multi-agent systems.

Palem is the Ken and Audrey Kennedy Professor of Computing at Rice University, where he is director and founder of the Nanyang Technological University-Rice Institute of Sustainable and Applied Infodynamics, with appointments in computer science and electrical and computer engineering. Palem is a leader in Embedded Systems research, having founded one of the earliest laboratories for research in academia dedicated to this field in 1994.