The New York Times http://nyti.ms/1ETbhDW

SCIENCE

Researchers Announce Advance in Image-Recognition Software

By JOHN MARKOFF NOV. 17, 2014

MOUNTAIN VIEW, Calif. — Two groups of scientists, working independently, have created artificial intelligence software capable of recognizing and describing the content of photographs and videos with far greater accuracy than ever before, sometimes even mimicking human levels of understanding.

Until now, so-called computer vision has largely been limited to recognizing individual objects. The new software, described on Monday by researchers at Google and at Stanford University, teaches itself to identify entire scenes: a group of young men playing Frisbee, for example, or a herd of elephants marching on a grassy plain.

The software then writes a caption in English describing the picture. Compared with human observations, the researchers found, the computer-written descriptions are surprisingly accurate.

The advances may make it possible to better catalog and search for the billions of images and hours of video available online, which are often poorly described and archived. At the moment, search engines like Google rely largely on written language accompanying an image or video to ascertain what it contains.

“I consider the pixel data in images and video to be the dark matter of the Internet,” said Fei-Fei Li, director of the Stanford Artificial Intelligence Laboratory, who led the research with Andrej Karpathy, a graduate student. “We are now starting to illuminate it.”

Dr. Li and Mr. Karpathy published their research as a Stanford University technical report. The Google team published their paper on arXiv.org, an open source site hosted by Cornell University.

In the longer term, the new research may lead to technology that helps the blind and robots navigate natural environments. But it also raises chilling possibilities for surveillance.

During the past 15 years, video cameras have been placed in a vast number of public and private spaces. In the future, the software operating the cameras will not only be able to identify particular humans via facial recognition, experts say, but also identify certain types of behavior, perhaps even automatically alerting authorities.

Two years ago Google researchers created image-recognition software and presented it with 10 million images taken from YouTube videos. Without human guidance, the program trained itself to recognize cats — a testament to the number of cat videos on YouTube.

Current artificial intelligence programs in new cars already can identify pedestrians and bicyclists from cameras positioned atop the windshield and can stop the car automatically if the driver does not take action to avoid a collision.

But “just single object recognition is not very beneficial,” said Ali Farhadi, a computer scientist at the University of Washington who has published research on software that generates sentences from digital pictures. “We’ve focused on objects, and we’ve ignored verbs,” he said, adding that these programs do not grasp what is going on in an image.

Both the Google and Stanford groups tackled the problem by refining software programs known as neural networks, inspired by our understanding of how the brain works. Neural networks can “train” themselves to discover similarities and patterns in data, even when their human creators do not know the patterns exist.

In living organisms, webs of neurons in the brain vastly outperform even the best computer-based networks in perception and pattern recognition. But by adopting some of the same architecture, computers are catching up, learning to identify patterns in speech and imagery with increasing accuracy. The advances are apparent to consumers who use Apple’s Siri personal assistant, for example, or Google’s image search.

Both groups of researchers employed similar approaches, weaving together two types of neural networks, one focused on recognizing images and the other on human language. In both cases the researchers trained the software with relatively small sets of digital images that had been annotated with descriptive sentences by humans.

After the software programs "learned" to see patterns in the pictures and description, the researchers turned them on previously unseen images. The programs were able to identify objects and actions with roughly double the accuracy of earlier efforts, although still nowhere near human perception capabilities.

"I was amazed that even with the small amount of training data that we were able to do so well," said Oriol Vinyals, a Google computer scientist who wrote the paper with Alexander Toshev, Samy Bengio and Dumitru Erhan, members of the Google Brain project. "The field is just starting, and we will see a lot of increases."

Computer vision specialists said that despite the improvements, these software systems had made only limited progress toward the goal of digitally duplicating human vision and, even more elusive, understanding.

"I don't know that I would say this is 'understanding' in the sense we want," said John R. Smith, a senior manager at I.B.M.'s T.J. Watson Research Center in Yorktown Heights, N.Y. "I think even the ability to generate language here is very limited."

But the Google and Stanford teams said that they expect to see significant increases in accuracy as they improve their software and train these programs with larger sets of annotated images. A research group led by Tamara L. Berg, a computer scientist at the University of North Carolina at Chapel Hill, is training a neural network with one million images annotated by humans.

"You're trying to tell the story behind the image," she said. "A natural scene will be very complex, and you want to pick out the most important objects in the image."

A version of this article appears in print on November 18, 2014, on page A13 of the New York edition with the headline: Advance Reported in Content-Recognition Software.