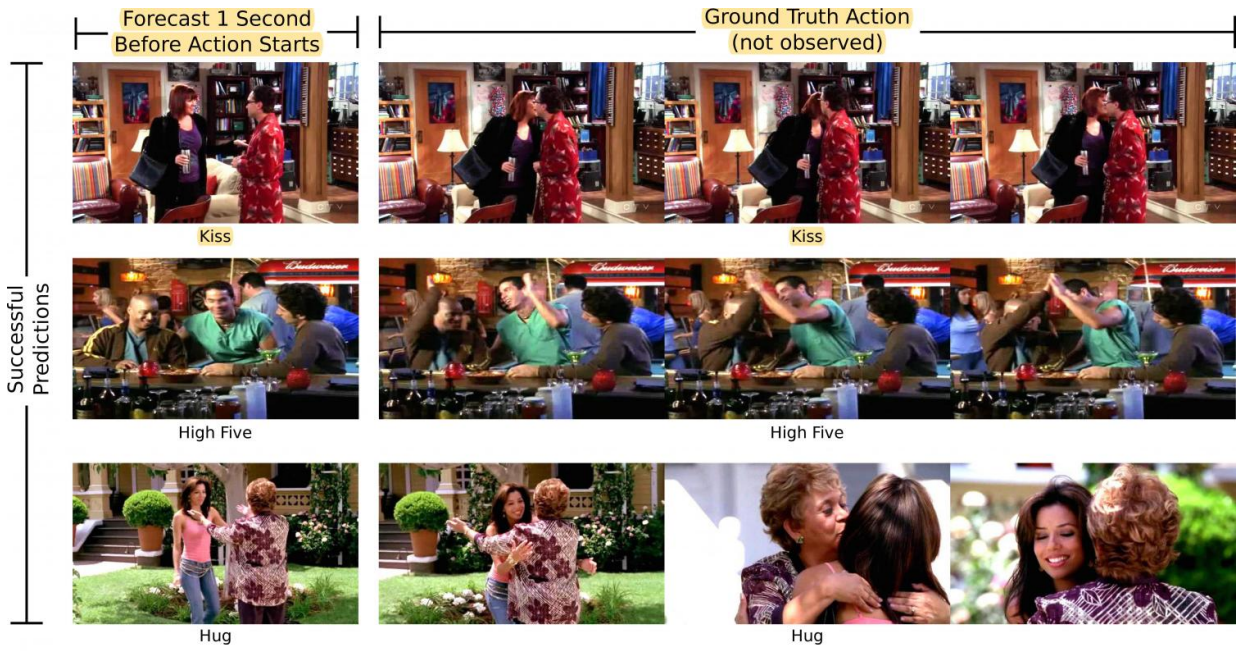
Deep-learning vision system anticipates human interactions using videos of TV shows

June 21 2016, by Adam Conner-Simons



The left-most column shows the frame before the action begins, with the algorithm's prediction below it. The right columns show the next frames of the video. Credit: Carl Vondrick/MIT CSAIL

When we see two people meet, we can often predict what happens next: a handshake, a hug, or maybe even a kiss. Our ability to anticipate actions is thanks to intuitions born out of a lifetime of experiences.

Machines, on the other hand, have trouble making use of complex

knowledge like that. Computer systems that predict actions would open up new possibilities ranging from robots that can better navigate human environments, to emergency response systems that predict falls, to Google Glass-style headsets that feed you suggestions for what to do in different situations.

This week researchers from MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL) have made an important new breakthrough in predictive vision, developing an algorithm that can anticipate interactions more accurately than ever before.

Trained on YouTube videos and TV shows such as "The Office" and "Desperate Housewives," the system can predict whether two individuals will hug, kiss, shake hands or slap five. In a second scenario, it could also anticipate what object is likely to appear in a video five seconds later.

While human greetings may seem like arbitrary actions to predict, the task served as a more easily controllable test case for the researchers to study.

"Humans automatically learn to anticipate actions through experience, which is what made us interested in trying to imbue computers with the same sort of common sense," says CSAIL PhD student Carl Vondrick, who is first author on a related paper that he will present this week at the International Conference on Computer Vision and Pattern Recognition (CVPR). "We wanted to show that just by watching large amounts of video, computers can gain enough knowledge to consistently make predictions about their surroundings."

Vondrick's co-authors include MIT Professor Antonio Torralba and former postdoc Hamed Pirsiavash, now a professor at the University of Maryland.

How it works

Past attempts at predictive computer-vision have generally taken one of two approaches.

The first method is to look at an image's individual pixels and use that knowledge to create a photorealistic "future" image, pixel by pixel—a task that Vondrick describes as "difficult for a professional painter, much less an algorithm." The second is to have humans label the scene for the computer in advance, which is impractical for being able to predict actions on a large scale.

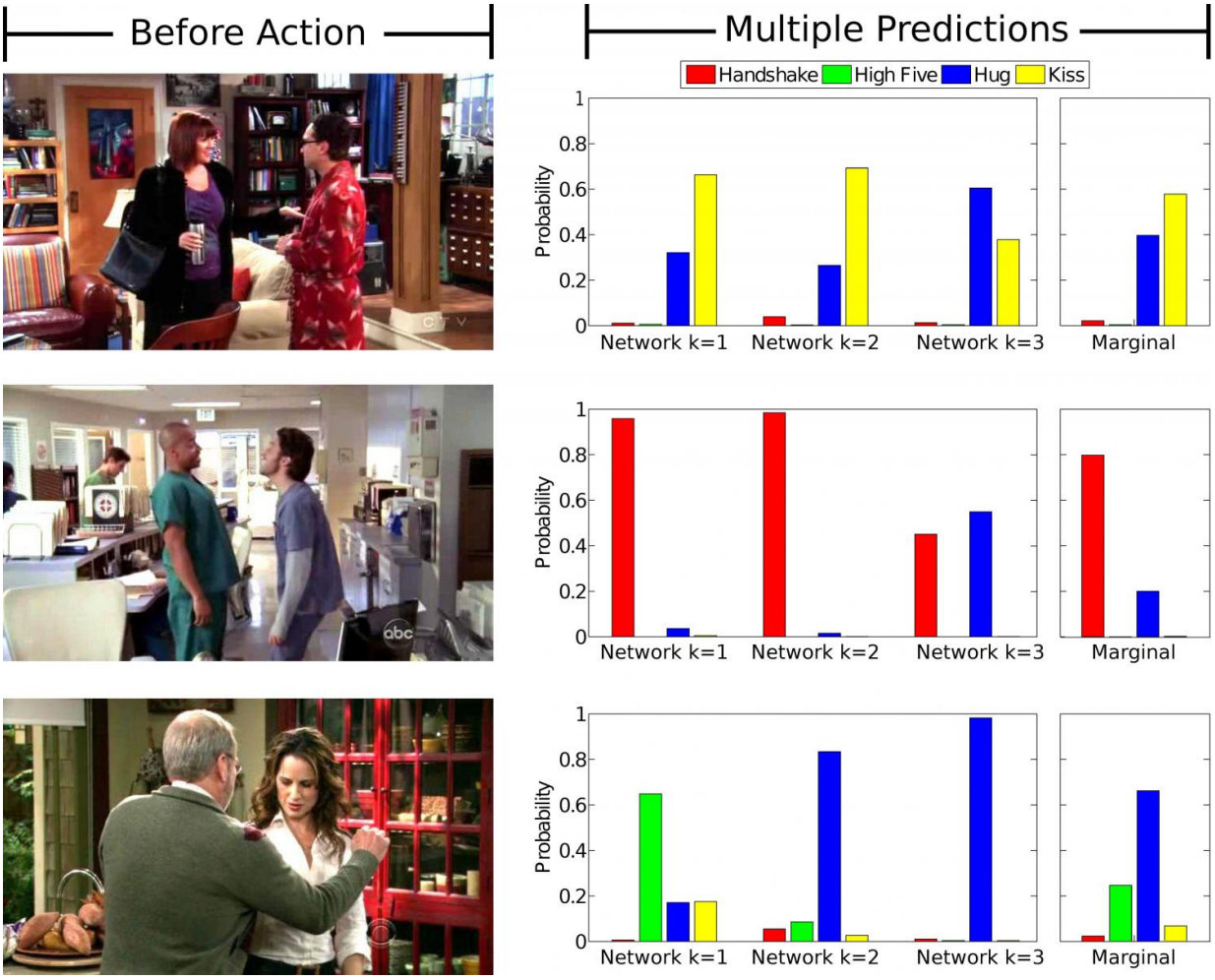
The CSAIL team instead created an algorithm that can predict "visual representations," which are basically freeze-frames showing different versions of what the scene might look like.

"Rather than saying that one pixel value is blue, the next one is red, and so on, visual representations reveal information about the larger image, such as a certain collection of pixels that represents a human face," Vondrick says.

The team's algorithm employs techniques from deep-learning, a field of artificial intelligence that uses systems called "neural networks" to teach computers to pore over massive amounts of data to find patterns on their own.

Each of the algorithm's networks predicts a representation is automatically classified as one of the four actions—in this case, a hug, handshake, high-five, or kiss. The system then merges those actions into one that it uses as its prediction. For example, three networks might predict a kiss, while another might use the fact that another person has entered the frame as a rationale for predicting a hug instead.

"A video isn't like a 'Choose Your Own Adventure' book where you can see all of the potential paths," says Vondrick. "The future is inherently ambiguous, so it's exciting to challenge ourselves to develop a system that uses these representations to anticipate all of the possibilities."



Given an input frame (left), the deep-learning model has multiple neural networks that each predict an action. The system then merges those actions into one that it uses as its prediction. Credit: Carl Vondrick/MIT CSAIL

How it did

After training the algorithm on 600 hours of unlabeled video, the team tested it on new videos showing both actions and objects.

When shown a video of people who are one second away from performing one of the four actions, the algorithm correctly predicted the action more than 43 percent of the time, which compares to existing algorithms that could only do 36 percent of the time.

In a second study, the algorithm was shown a frame from a video and asked to predict what object will appear five seconds later. For example, seeing someone open a microwave might suggest the future presence of a coffee mug. The algorithm predicted the object in the frame 30 percent more accurately than baseline measures, though the researchers caution that it still only has an average precision of 11 percent.

It's worth noting that even humans make mistakes on these tasks: for example, human subjects were only able to correctly predict the action 71 percent of the time.

"There's a lot of subtlety to understanding and forecasting human interactions," says Vondrick. "We hope to be able to work off of this example to be able to soon predict even more complex tasks."

Looking forward

While the algorithms aren't yet accurate enough for practical applications, Vondrick says that future versions could be used for everything from robots that develop better action plans to security cameras that can alert emergency responders when someone who has fallen or gotten injured.

"I'm excited to see how much better the algorithms get if we can feed them a lifetime's worth of videos," says Vondrick. "We might see some significant improvements that would get us closer to using predictive-vision in real-world situations."

The work was supported by a grant from the National Science Foundation, along with a Google faculty research award for Torralba and a Google PhD fellowship for Vondrick.

More information: Paper: "Anticipating Visual Representations with Unlabeled Videos" web.mit.edu/vondrick/prediction.pdf

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