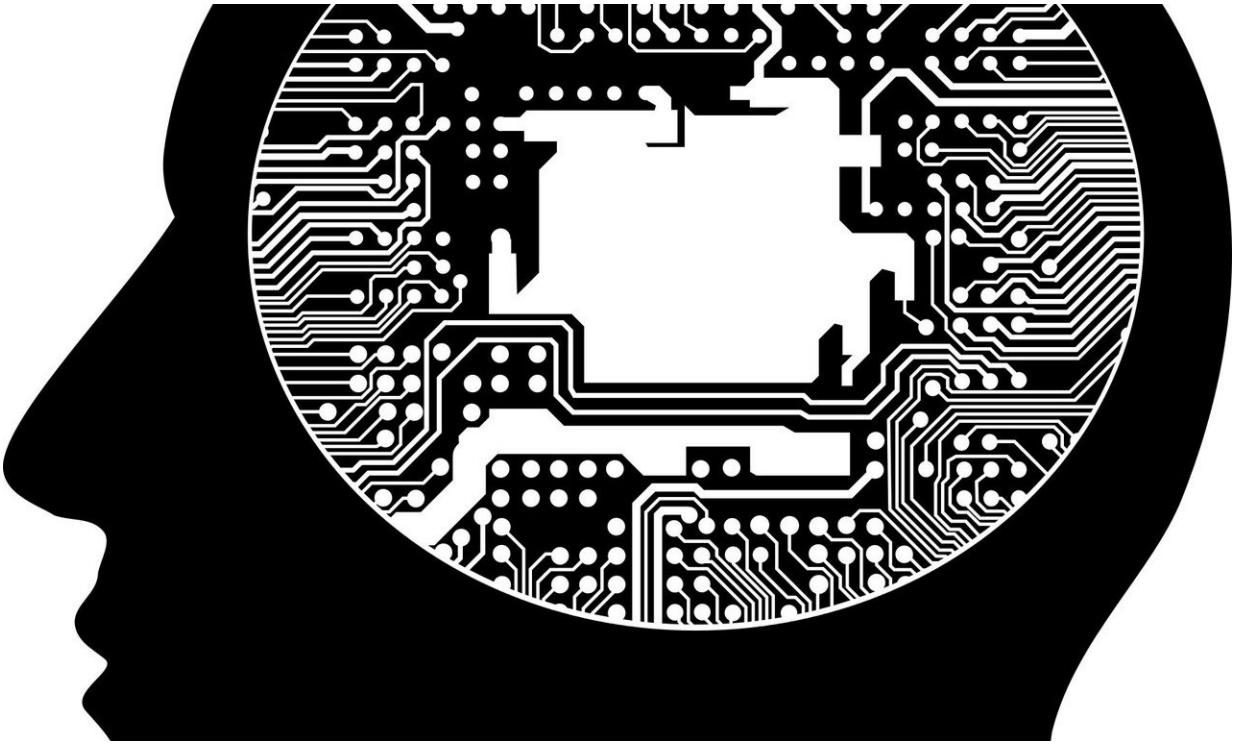
Helping computers fill in the gaps between video frames

September 13 2018, by Rob Matheson



Credit: CC0 Public Domain

Given only a few frames of a video, humans can usually surmise what is happening and will happen on screen. If we see an early frame of stacked cans, a middle frame with a finger at the stack's base, and a late frame showing the cans toppled over, we can guess that the finger knocked down the cans. Computers, however, struggle with this concept.

In a paper being presented at this week's European Conference on Computer Vision, MIT researchers describe an add-on module that helps artificial intelligence systems called convolutional neural networks, or CNNs, to fill in the gaps between video frames to greatly improve the network's activity recognition.

The researchers' module, called Temporal Relation Network (TRN), learns how objects change in a video at different times. It does so by analyzing a few key frames depicting an activity at different stages of the video—such as stacked objects that are then knocked down. Using the same process, it can then recognize the same type of activity in a new video.

In experiments, the module outperformed existing models by a large margin in recognizing hundreds of basic activities, such as poking objects to make them fall, tossing something in the air, and giving a thumbs-up. It also more accurately predicted what will happen next in a video—showing, for example, two hands making a small tear in a sheet of paper—given only a small number of early frames.

One day, the module could be used to help robots better understand what's going on around them.

"We built an artificial intelligence system to recognize the transformation of objects, rather than appearance of objects," says Bolei Zhou, a former Ph.D. student in the Computer Science and Artificial Intelligence Laboratory (CSAIL) who is now an assistant professor of computer science at the Chinese University of Hong Kong. "The system doesn't go through all the frames—it picks up key frames and, using the temporal relation of frames, recognize what's going on. That improves the efficiency of the system and makes it run in real-time accurately."

Co-authors on the paper are CSAIL principal investigator Antonio

Torralba, who is also a professor in the Department of Electrical Engineering and Computer Science; CSAIL Principal Research Scientist Aude Oliva; and CSAIL Research Assistant Alex Andonian.

Picking up key frames

Two common CNN modules being used for activity recognition today suffer from efficiency and accuracy drawbacks. One model is accurate but must analyze each video frame before making a prediction, which is computationally expensive and slow. The other type, called two-stream network, is less accurate but more efficient. It uses one stream to extract features of one video frame, and then merges the results with "optical flows," a stream of extracted information about the movement of each pixel. Optical flows are also computationally expensive to extract, so the model still isn't that efficient.

"We wanted something that works in between those two models—getting efficiency and accuracy," Zhou says.

The researchers trained and tested their module on three crowdsourced datasets of short videos of various performed activities. The first dataset, called Something-Something, built by the company TwentyBN, has more than 200,000 videos in 174 action categories, such as poking an object so it falls over or lifting an object. The second dataset, Jester, contains nearly 150,000 videos with 27 different hand gestures, such as giving a thumbs-up or swiping left. The third, Charades, built by Carnegie Mellon University researchers, has nearly 10,000 videos of 157 categorized activities, such as carrying a bike or playing basketball.

When given a video file, the researchers' module simultaneously processes ordered frames—in groups of two, three, and four—spaced some time apart. Then it quickly assigns a probably that the object's transformation across those frames matches a specific activity class. For

instance, if it processes two frames, where the later frame shows an object at the bottom of the screen and the earlier shows the object at the top, it will assign a high probability to the activity class, "moving object down." If a third frame shows the object in the middle of the screen, that probability increases even more, and so on. From this, it learns object-transformation features in frames that most represent a certain class of activity.

Recognizing and forecasting activities

In testing, a CNN equipped with the new module accurately recognized many activities using two frames, but the accuracy increased by sampling more frames. For Jester, the module achieved top accuracy of 95 percent in activity recognition, beating out several existing models.

It even guessed right on ambiguous classifications: Something-Something, for instance, included actions such as "pretending to open a book" versus "opening a book." To discern between the two, the module just sampled a few more key frames, which revealed, for instance, a hand near a book in an early frame, then on the book, then moved away from the book in a later frame.

Some other activity-recognition models also process key frames but don't consider temporal relationships in frames, which reduces their accuracy. The researchers report that their TRN module nearly doubles in accuracy over those key-frame models in certain tests.

The module also outperformed models on forecasting an activity, given limited frames. After processing the first 25 percent of frames, the module achieved accuracy several percentage points higher than a baseline model. With 50 percent of the frames, it achieved 10 to 40 percent higher accuracy. Examples include determining that a paper would be torn just a little, based how two hands are positioned on the

paper in early frames, and predicting that a raised hand, shown facing forward, would swipe down.

"That's important for robotics applications," Zhou says. "You want [a robot] to anticipate and forecast what will happen early on, when you do a specific action."

Next, the researchers aim to improve the module's sophistication. First step is implementing object recognition together with activity recognition. Then, they hope to add in "intuitive physics," meaning helping it understand real-world physical properties of objects. "Because we know a lot of the physics inside these videos, we can train module to learn such physics laws and use those in recognizing new videos," Zhou says. "We also open source all the code and models. Activity understanding is an exciting area of artificial intelligence right now."

More information: "Temporal Relational Reasoning in Videos" arXiv:1711.08496 [cs.CV] arxiv.org/abs/1711.08496

Provided by Massachusetts Institute of Technology