Object and scene recognition software work together to understand video content

June 23 2016

Researchers from Disney Research and Shanghai's Fudan University have used deep learning techniques to train computer software to recognize events in videos, even categories of events that the software has not previously seen.

Their approach uses both scene and object features from the video and enables associations between these visual elements and each type of event to be automatically determined and weighted by a machine-learning architecture known as a neural network.

Notably, this approach not only works better than other methods in recognizing events in videos, but is significantly better at identifying events that that the computer program has never or rarely encountered previously, said Leonid Sigal, senior research scientist at Disney Research. These events can include such things as riding a horse, baking cookies or eating at a restaurant.

"Automated techniques are essential for indexing, searching and analyzing the incredible amount of video being created and uploaded daily to the Internet," said Jessica Hodgins, vice president at Disney Research. "With multiple hours of video being uploaded to YouTube every second, there's no way to describe all of that content manually. And if we don't know what's in all those videos, we can't find things we need and much of the videos' potential value is lost."

The researchers will present their method June 26 at the Computer

Vision and Pattern Recognition (CVPR 2016) conference in Las Vegas. In addition to Sigal, the team includes Yanwei Fu, a post-doctoral researcher at Disney Research; Zuxuan Wu, a graduate student in computer science at Fudan who conducted this research during his internship at Disney Research and Yu-Gang Jiang, a professor of computer science at Fudan.

Understanding the content of a video, particularly user-generated video, is a difficult challenge for computer vision because video content can vary so much. Even when the content - a particular concert, for instance - is the same, it can look very different depending on the perspective from which it was shot and on lighting conditions.

Computer vision researchers have had some success using a deep learning approach involving Convolutional Neural Networks (CNNs) to identify events when a large amount of labeled examples are available to train the computer model. But that method doesn't work if few labeled examples are available to train the model, so scaling it up to include thousands, if not tens of thousands, of additional classes of events would be difficult.

Wu said the team's approach enables the computer model to identify objects and scenes associated with each activity or event and figure out how much weight to give each. "Eating sushi," for instance, might be associated with chopsticks, a restaurant setting, an indoor setting and serving dishes.

"The model learns what is important for each event," he said.

When presented with an event that it has not previously encountered, the model can identify objects and scenes that it already has associated with similar events to help it classify the new event, Sigal said. If it already is familiar with "eating pasta" and "eating rice," for instance, it might

reason that elements associated with one or the other - chopsticks, bowls, restaurant settings, etc.,—might be associated with "eating Soba or Udon noodles."

This ability to extend its knowledge into events not previously seen, or for which labeled examples are limited, makes it possible to scale up the model to include an ever-increasing number of event classes, he added.

Provided by Disney Research

Citation: Object and scene recognition software work together to understand video content (2016, June 23) retrieved 23 April 2024 from https://phys.org/news/2016-06-scene-recognition-software-video-content.html