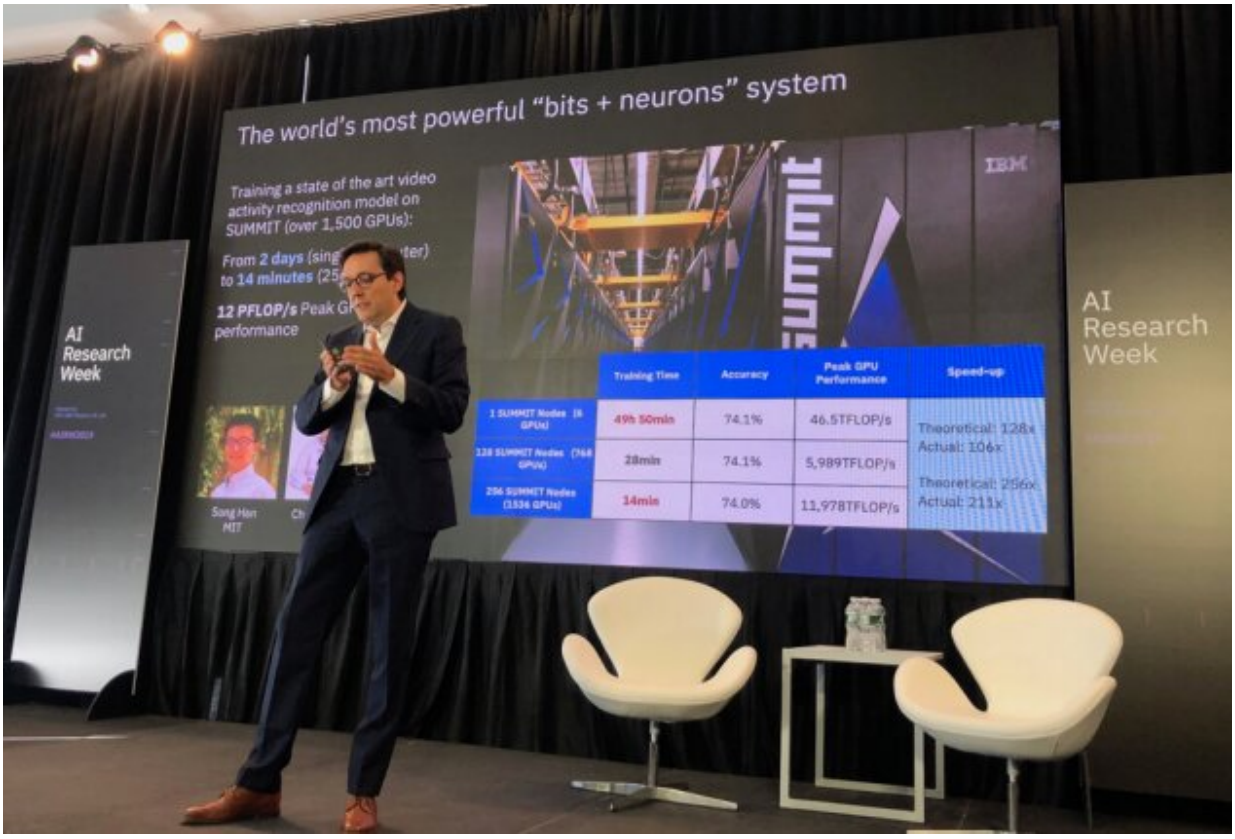
Faster video recognition for the smartphone era

October 14 2019, by Kim Martineau



A new technique for training video recognition models is up to three times faster than current state-of-the-art methods while improving runtime performance on mobile devices. The work was recently highlighted by Dario Gil (above), director of IBM Research, at the MIT-IBM Watson AI Lab's AI Research Week in Cambridge, Massachusetts. Photo: Song Han

A branch of machine learning called deep learning has helped computers surpass humans at well-defined visual tasks like reading medical scans, but as the technology expands into interpreting videos and real-world events, the models are getting larger and more computationally intensive.

By one estimate, training a video-recognition model can take up to 50 times more data and eight times more processing power than training an image-classification model. That's a problem as demand for processing power to train deep learning models continues to rise exponentially and concerns about AI's massive carbon footprint grow. Running large video-recognition models on low-power mobile devices, where many AI applications are heading, also remains a challenge.

Song Han, an assistant professor at MIT's Department of Electrical Engineering and Computer Science (EECS), is tackling the problem by designing more efficient deep learning models. In a paper at the International Conference on Computer Vision, Han, MIT graduate student Ji Lin and MIT-IBM Watson AI Lab researcher Chuang Gan, outline a method for shrinking video-recognition models to speed up training and improve runtime performance on smartphones and other mobile devices. Their method makes it possible to shrink the model to one-sixth the size by reducing the 150 million parameters in a state-of-the-art model to 25 million parameters.

"Our goal is to make AI accessible to anyone with a low-power device," says Han. "To do that, we need to design efficient AI models that use less energy and can run smoothly on edge devices, where so much of AI is moving."

The falling cost of cameras and video-editing software and the rise of new video-streaming platforms has flooded the internet with new content. Each hour, 30,000 hours of new video are uploaded to YouTube alone. Tools to catalog that content more efficiently would help viewers

and advertisers locate videos faster, the researchers say. Such tools would also help institutions like hospitals and nursing homes to run AI applications locally, rather than in the cloud, to keep sensitive data private and secure.

Underlying image and video-recognition models are neural networks, which are loosely modeled on how the brain processes information. Whether it's a digital photo or sequence of video images, neural nets look for patterns in the pixels and build an increasingly abstract representation of what they see. With enough examples, neural nets "learn" to recognize people, objects, and how they relate.

Top video-recognition models currently use three-dimensional convolutions to encode the passage of time in a sequence of images, which creates bigger, more computationally-intensive models. To reduce the calculations involved, Han and his colleagues designed an operation they call a temporal shift module which shifts the feature maps of a selected video frame to its neighboring frames. By mingling spatial representations of the past, present, and future, the model gets a sense of time passing without explicitly representing it.

The result: a model that outperformed its peers at recognizing actions in the Something-Something video dataset, earning first place in version 1 and version 2, in recent public rankings. An online version of the shift module is also nimble enough to read movements in real-time. In a recent demo, Lin, a Ph.D. student in EECS, showed how a single-board computer rigged to a video camera could instantly classify hand gestures with the amount of energy to power a bike light.

Normally it would take about two days to train such a powerful model on a machine with just one graphics processor. But the researchers managed to borrow time on the U.S. Department of Energy's Summit supercomputer, currently ranked the fastest on Earth. With Summit's

extra firepower, the researchers showed that with 1,536 graphics processors the model could be trained in just 14 minutes, near its theoretical limit. That's up to three times faster than 3-D state-of-the-art models, they say.

Dario Gil, director of IBM Research, highlighted the work in his recent opening remarks at AI Research Week hosted by the MIT-IBM Watson AI Lab.

"Compute requirements for large AI training jobs is doubling every 3.5 months," he said later. "Our ability to continue pushing the limits of the technology will depend on strategies like this that match hyper-efficient algorithms with powerful machines."

More information: TSM: Temporal Shift Module for Efficient Video Understanding: arXiv:1811.08383v3 [cs.CV]: arxiv.org/pdf/1811.08383.pdf

Training Kinetics in 15 Minutes: Large-scale Distributed Training on Videos: arXiv:1910.00932v1 [cs.CV]: arxiv.org/pdf/1910.00932.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