Training computers to recognize dynamic events

April 5 2018, by Meg Murphy

Aude Oliva (right), a principal research scientist at the Computer Science and Artificial Intelligence Laboratory and Dan Gutfreund (left), a principal investigator at the MIT–IBM Watson AI Laboratory and a staff member at IBM Research, are the principal investigators for the Moments in Time Dataset, one of the projects related to AI algorithms funded by the MIT–IBM Watson AI Laboratory. Credit: John Mottern/Feature Photo Service for IBM

A person watching videos that show things opening—a door, a book, curtains, a blooming flower, a yawning dog—easily understands the same type of action is depicted in each clip.

"Computer models fail miserably to identify these things. How do humans do it so effortlessly?" asks Dan Gutfreund, a principal investigator at the MIT-IBM Watson AI Laboratory and a staff member at IBM Research. "We process information as it happens in space and time. How can we teach computer models to do that?"

Such are the big questions behind one of the new projects underway at the MIT-IBM Watson AI Laboratory, a collaboration for research on the frontiers of artificial intelligence. Launched last fall, the lab connects MIT and IBM researchers together to work on AI algorithms, the application of AI to industries, the physics of AI, and ways to use AI to advance shared prosperity.

The Moments in Time dataset is one of the projects related to AI algorithms that is funded by the lab. It pairs Gutfreund with Aude Oliva, a principal research scientist at the MIT Computer Science and Artificial Intelligence Laboratory, as the project's principal investigators. Moments in Time is built on a collection of 1 million annotated videos of dynamic events unfolding within three seconds. Gutfreund and Oliva, who is also the MIT executive director at the MIT-IBM Watson AI Lab, are using these clips to address one of the next big steps for AI: teaching machines to recognize actions.

Learning from dynamic scenes

The goal is to provide deep-learning algorithms with large coverage of an ecosystem of visual and auditory moments that may enable models to learn information that isn't necessarily taught in a supervised manner and to generalize to novel situations and tasks, say the researchers.

"As we grow up, we look around, we see people and objects moving, we hear sounds that people and object make. We have a lot of visual and auditory experiences. An AI system needs to learn the same way and be fed with videos and dynamic information," Oliva says.

For every action category in the dataset, such as cooking, running, or opening, there are more than 2,000 videos. The short clips enable computer models to better learn the diversity of meaning around specific actions and events.

"This dataset can serve as a new challenge to develop AI models that scale to the level of complexity and abstract reasoning that a human processes on a daily basis," Oliva adds, describing the factors involved. Events can include people, objects, animals, and nature. They may be symmetrical in time—for example, opening means closing in reverse order. And they can be transient or sustained.

Oliva and Gutfreund, along with additional researchers from MIT and IBM, met weekly for more than a year to tackle technical issues, such as how to choose the action categories for annotations, where to find the videos, and how to put together a wide array so the AI system learns without bias. The team also developed machine-learning models, which were then used to scale the data collection. "We aligned very well because we have the same enthusiasm and the same goal," says Oliva.

Augmenting human intelligence

One key goal at the lab is the development of AI systems that move beyond specialized tasks to tackle more complex problems and benefit from robust and continuous learning. "We are seeking new algorithms that not only leverage big data when available, but also learn from limited data to augment human intelligence," says Sophie V. Vandebroek, chief operating officer of IBM Research, about the

collaboration.

In addition to pairing the unique technical and scientific strengths of each organization, IBM is also bringing MIT researchers an influx of resources, signaled by its $240 million investment in AI efforts over the next 10 years, dedicated to the MIT-IBM Watson AI Lab. And the alignment of MIT-IBM interest in AI is proving beneficial, according to Oliva.

"IBM came to MIT with an interest in developing new ideas for an artificial intelligence system based on vision. I proposed a project where we build data sets to feed the model about the world. It had not been done before at this level. It was a novel undertaking. Now we have reached the milestone of 1 million videos for visual AI training, and people can go to our website, download the dataset and our deep-learning computer models, which have been taught to recognize actions."

Qualitative results so far have shown models can recognize moments well when the action is well-framed and close up, but they misfire when the category is fine-grained or there is background clutter, among other things. Oliva says that MIT and IBM researchers have submitted an article describing the performance of neural network models trained on the dataset, which itself was deepened by shared viewpoints. "IBM researchers gave us ideas to add action categories to have more richness in areas like health care and sports. They broadened our view. They gave us ideas about how AI can make an impact from the perspective of business and the needs of the world," she says.

This first version of the Moments in Time dataset is one of the largest human-annotated video datasets capturing visual and audible short events, all of which are tagged with an action or activity label among 339 different classes that include a wide range of common verbs. The researchers intend to produce more datasets with a variety of levels of

abstraction to serve as stepping stones toward the development of learning algorithms that can build analogies between things, imagine and synthesize novel events, and interpret scenarios.

In other words, they are just getting started, says Gutfreund. "We expect the Moments in Time dataset to enable models to richly understand actions and dynamics in videos."

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