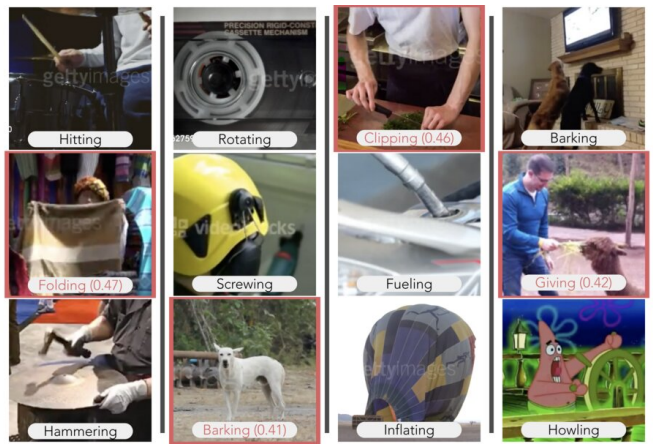
Toward a machine learning model that can reason about everyday actions

1 September 2020, by Kim Martineau



A computer vision model developed by researchers at MIT, IBM, and Columbia University can compare and contrast dynamic events captured on video to tease out the high-level concepts connecting them. In a set of experiments, the model picked out the video in each vertical-column set that conceptually didn't belong. Highlighted in red, the odd-one-out videos show a woman folding a blanket, a dog barking, a man chopping greens, and a man offering grass to a llama. Credit: Allen Lee

The ability to reason abstractly about events as they unfold is a defining feature of human intelligence. We know instinctively that crying and writing are means of communicating, and that a panda falling from a tree and a plane landing are variations on descending.

Organizing the world into abstract categories does not come easily to computers, but in recent years researchers have inched closer by training machine learning models on words and images infused with structural information about the world, and how objects, animals, and actions relate. In a new study at the European Conference on Computer Vision this month, researchers unveiled a hybrid language-vision model that can compare and contrast a set of dynamic events captured on video to tease out the high-level concepts connecting them.

Their model did as well as or better than humans at two types of visual reasoning tasks—picking the video that conceptually best completes the set, and picking the video that doesn't fit. Shown videos of a dog barking and a man howling beside his dog, for example, the model completed the set by picking the crying baby from a set of five videos. Researchers replicated their results on two datasets for training AI systems in action recognition: MIT's Multi-Moments in Time and DeepMind's Kinetics.

"We show that you can build abstraction into an AI system to perform ordinary visual reasoning tasks close to a human level," says the study's senior author Aude Oliva, a senior research scientist at MIT, co-director of the MIT Quest for Intelligence, and MIT director of the MIT-IBM Watson AI Lab. "A model that can recognize abstract events will give more accurate, logical predictions and be more useful for decision-making."

As deep neural networks become expert at recognizing objects and actions in photos and video, researchers have set their sights on the next milestone: abstraction, and training models to reason about what they see. In one approach, researchers have merged the pattern-matching power of deep nets with the logic of symbolic programs to teach a model to interpret complex object relationships in a scene. Here, in another approach, researchers capitalize on the relationships embedded in the meanings of words to give their model visual reasoning power.

"Language representations allow us to integrate contextual information learned from text databases into our visual models," says study co-author Mathew Monfort, a research scientist at MIT's Computer Science and Artificial Intelligence

Laboratory (CSAIL). "Words like 'running,' 'lifting,' and 'boxing' share some common characteristics that make them more closely related to the concept 'exercising,' for example, than 'driving.' "

Using WordNet, a database of word meanings, the researchers mapped the relation of each action-class label in Moments and Kinetics to the other labels in both datasets. Words like "sculpting," "carving," and "cutting," for example, were connected to higher-level concepts like "crafting," "making art," and "cooking." Now when the model recognizes an activity like sculpting, it can pick out conceptually similar activities in the dataset.

This relational graph of abstract classes is used to train the model to perform two basic tasks. Given a set of videos, the model creates a numerical representation for each video that aligns with the word representations of the actions shown in the video. An abstraction module then combines the representations generated for each video in the set to create a new set representation that is used to identify the abstraction shared by all the videos in the set.

To see how the model would do compared to humans, the researchers asked human subjects to perform the same set of visual reasoning tasks online. To their surprise, the model performed as well as humans in many scenarios, sometimes with unexpected results. In a variation on the set completion task, after watching a video of someone wrapping a gift and covering an item in tape, the model suggested a video of someone at the beach burying someone else in the sand.

"It's effectively 'covering,' but very different from the visual features of the other clips," says Camilo Fosco, a Ph.D. student at MIT who is co-first author of the study with Ph.D. student Alex Andonian. "Conceptually it fits, but I had to think about it."

Limitations of the model include a tendency to overemphasize some features. In one case, it suggested completing a set of sports videos with a video of a baby and a ball, apparently associating balls with exercise and competition.

A deep learning model that can be trained to "think" more abstractly may be capable of learning with fewer data, say researchers. Abstraction also paves the way toward higher-level, more human-like reasoning.

"One hallmark of human cognition is our ability to describe something in relation to something else—to compare and to contrast," says Oliva. "It's a rich and efficient way to learn that could eventually lead to machine learning models that can understand analogies and are that much closer to communicating intelligently with us."

More information: We Have So Much In Common: Modeling Semantic Relational Set Abstractions in Videos: abstraction.csail.mit.edu/stat …a_Ready.bc714a4e.pdf

Monfort et al., Multi-Moments in Time: Learning and Interpreting Models for Multi-Action Video Understanding. arXiv:1911.00232 [cs.CV]. arxiv.org/abs/1911.00232v2