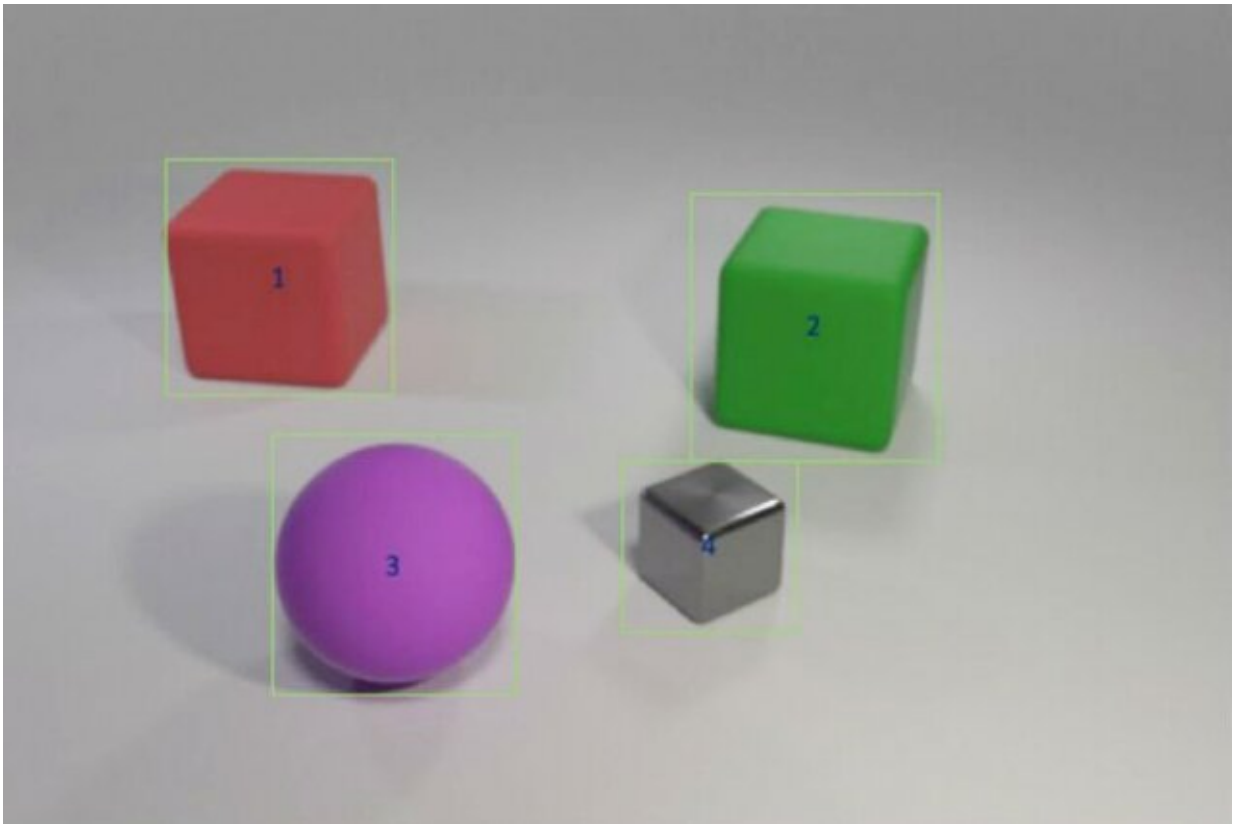
Teaching machines to reason about what they see

April 3 2019, by Kim Martineau



Researchers trained a hybrid AI model to answer questions like “Does the red object left of the green cube have the same shape as the purple matte thing?” by feeding it examples of object colors and shapes followed by more complex scenarios involving multi-object comparisons. The model could transfer this knowledge to new scenarios as well as or better than state-of-the-art models using a fraction of the training data. Credit: Justin Johnson

A child who has never seen a pink elephant can still describe one—unlike a computer. "The computer learns from data," says Jiajun Wu, a Ph.D. student at MIT. "The ability to generalize and recognize something you've never seen before—a pink elephant—is very hard for machines."

Deep learning systems interpret the world by picking out statistical patterns in data. This form of machine learning is now everywhere, automatically tagging friends on Facebook, narrating Alexa's latest weather forecast, and delivering fun facts via Google search. But statistical learning has its limits. It requires tons of data, has trouble explaining its decisions, and is terrible at applying past knowledge to new situations; It can't comprehend an elephant that's pink instead of gray.

To give computers the ability to reason more like us, artificial intelligence (AI) researchers are returning to abstract, or symbolic, programming. Popular in the 1950s and 1960s, symbolic AI wires in the rules and logic that allow machines to make comparisons and interpret how objects and entities relate. Symbolic AI uses less data, records the chain of steps it takes to reach a decision, and when combined with the brute processing power of statistical neural networks, it can even beat humans in a complicated image comprehension test.

A new study by a team of researchers at MIT, MIT-IBM Watson AI Lab, and DeepMind shows the promise of merging statistical and symbolic AI. Led by Wu and Joshua Tenenbaum, a professor in MIT's Department of Brain and Cognitive Sciences and the Computer Science and Artificial Intelligence Laboratory, the team shows that its hybrid model can learn object-related concepts like color and shape, and leverage that knowledge to interpret complex object relationships in a scene. With minimal training data and no explicit programming, their model could transfer concepts to larger scenes and answer increasingly

tricky questions as well as or better than its state-of-the-art peers. The team presents its results at the International Conference on Learning Representations in May.

"One way children learn concepts is by connecting words with images," says the study's lead author Jiayuan Mao, an undergraduate at Tsinghua University who worked on the project as a visiting fellow at MIT. "A machine that can learn the same way needs much less data, and is better able to transfer its knowledge to new scenarios."

The study is a strong argument for moving back toward abstract-program approaches, says Jacob Andreas, a recent graduate of the University of California at Berkeley, who starts at MIT as an assistant professor this fall and was not involved in the work. "The trick, it turns out, is to add more symbolic structure, and to feed the neural networks a representation of the world that's divided into objects and properties rather than feeding it raw images," he says. "This work gives us insight into what machines need to understand before language learning is possible."

The team trained their model on images paired with related questions and answers, part of the CLEVR image comprehension test developed at Stanford University. As the model learns, the questions grow progressively harder, from, "What's the color of the object?" to "How many objects are both right of the green cylinder and have the same material as the small blue ball?" Once object-level concepts are mastered, the model advances to learning how to relate objects and their properties to each other.

Like other hybrid AI models, MIT's works by splitting up the task. A perception module of neural networks crunches the pixels in each image and maps the objects. A language module, also made of neural nets, extracts a meaning from the words in each sentence and creates symbolic

programs, or instructions, that tell the machine how to answer the question. A third reasoning module runs the symbolic programs on the scene and gives an answer, updating the model when it makes mistakes.

Key to the team's approach is a perception module that translates the image into an object-based representation, making the programs easier to execute. Also unique is what they call curriculum learning, or selectively training the model on concepts and scenes that grow progressively more difficult. It turns out that feeding the machine data in a logical way, rather than haphazardly, helps the model learn faster while improving accuracy.

Once the model has a solid foundation, it can interpret new scenes and concepts, and increasingly difficult questions, almost perfectly. Asked to answer an unfamiliar question like, "What's the shape of the big yellow thing?" it outperformed its peers at Stanford and nearby MIT Lincoln Laboratory with a fraction of the data.

While other models trained on the full CLEVR dataset of 70,000 images and 700,000 questions, the MIT-IBM model used 5,000 images and 100,000 questions. As the model built on previously learned concepts, it absorbed the programs underlying each question, speeding up the training process.

Though statistical, deep learning models are now embedded in daily life, much of their decision process remains hidden from view. This lack of transparency makes it difficult to anticipate where the system is susceptible to manipulation, error, or bias. Adding a symbolic layer can open the black box, explaining the growing interest in hybrid AI systems.

"Splitting the task up and letting programs do some of the work is the key to building interpretability into deep learning models," says Lincoln Laboratory researcher David Mascharka, whose hybrid model,

Transparency by Design Network, is benchmarked in the MIT-IBM study.

The MIT-IBM team is now working to improve the model's performance on real-world photos and extending it to video understanding and robotic manipulation. Other authors of the study are Chuang Gan and Pushmeet Kohli, researchers at the MIT-IBM Watson AI Lab and DeepMind, respectively.

More information: The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision. openreview.net/pdf?id=rJgMlhRctm

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