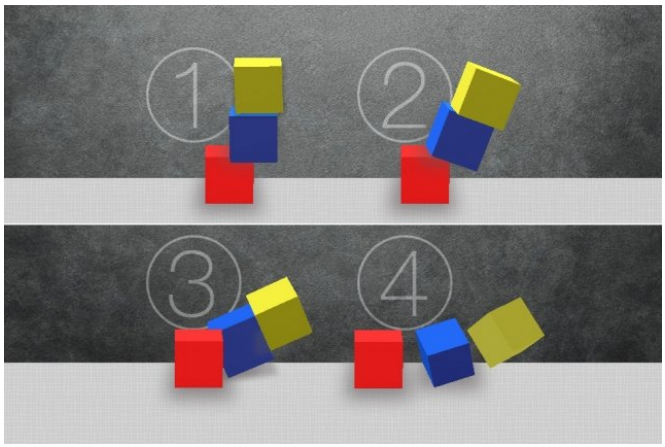
Computer systems predict objects' responses to physical forces

14 December 2017, by Larry Hardesty



As part of an investigation into the nature of humans' physical intuitions, MIT researchers trained a neural network to predict how unstably stacked blocks would respond to the force of gravity. Credit: Christine Daniloff/MIT

Josh Tenenbaum, a professor of brain and cognitive sciences at MIT, directs research on the development of intelligence at the Center for Brains, Minds, and Machines, a multi-university, multidisciplinary project based at MIT that seeks to explain and replicate human intelligence.

Presenting their work at this year's Conference on Neural Information Processing Systems, Tenenbaum and one of his students, Jiajun Wu, are co-authors on four papers that examine the fundamental cognitive abilities that an intelligent agent requires to navigate the world: discerning distinct objects and inferring how they respond to physical forces.

By building computer systems that begin to approximate these capacities, the researchers believe they can help answer questions about what information-processing resources human beings use at what stages of development. Along the way, the researchers might also generate some insights useful for robotic vision systems.

"The common theme here is really learning to perceive physics," Tenenbaum says. "That starts with seeing the full 3-D shapes of objects, and multiple objects in a scene, along with their physical properties, like mass and friction, then reasoning about how these objects will move over time. Jiajun's four papers address this whole space. Taken together, we're starting to be able to build machines that capture more and more of people's basic understanding of the physical world."

Three of the papers deal with inferring information about the physical structure of objects, from both visual and aural data. The fourth deals with predicting how objects will behave on the basis of that data.

Two-way street

Something else that unites all four papers is their unusual approach to machine learning, a technique in which computers learn to perform computational tasks by analyzing huge sets of training data. In a typical machine-learning system, the training data are labeled: Human analysts will have, say, identified the objects in a visual scene or transcribed the words of a spoken sentence. The system attempts to learn what features of the data correlate with what labels, and it's judged on how well it labels previously unseen data.

In Wu and Tenenbaum's new papers, the system is trained to infer a physical model of the world—the 3-D shapes of objects that are mostly hidden from view, for instance. But then it works backward, using the model to resynthesize the input data, and its performance is judged on how well the reconstructed data matches the original data.

For instance, using visual images to build a 3-D model of an object in a scene requires stripping

away any occluding objects; filtering out confounding visual textures, reflections, and shadows; and inferring the shape of unseen surfaces. Once Wu and Tenenbaum's system has built such a model, however, it rotates it in space and adds visual textures back in until it can approximate the input data.

Indeed, two of the researchers' four papers address the complex problem of inferring 3-D models from visual data. On those papers, they're joined by four other MIT researchers, including William Freeman, the Perkins Professor of Electrical Engineering and Computer Science, and by colleagues at DeepMind, ShanghaiTech University, and Shanghai Jiao Tong University.

Divide and conquer

The researchers' system is based on the influential theories of the MIT neuroscientist David Marr, who died in 1980 at the tragically young age of 35. Marr hypothesized that in interpreting a visual scene, the brain first creates what he called a 2.5-D sketch of the objects it contained—a representation of just those surfaces of the objects facing the viewer. Then, on the basis of the 2.5-D sketch—not the raw visual information about the scene—the brain infers the full, three-dimensional shapes of the objects.

"Both problems are very hard, but there's a nice way to disentangle them," Wu says. "You can do them one at a time, so you don't have to deal with both of them at the same time, which is even harder."

Wu and his colleagues' system needs to be trained on data that include both visual images and 3-D models of the objects the images depict. Constructing accurate 3-D models of the objects depicted in real photographs would be prohibitively time consuming, so initially, the researchers train their system using synthetic data, in which the visual image is generated from the 3-D model, rather than vice versa. The process of creating the data is like that of creating a computer-animated film.

Once the system has been trained on synthetic data, however, it can be fine-tuned using real data. That's because its ultimate performance criterion is the accuracy with which it reconstructs the input data. It's still building 3-D models, but they don't need to be compared to human-constructed models for performance assessment.

In evaluating their system, the researchers used a measure called intersection over union, which is common in the field. On that measure, their system outperforms its predecessors. But a given intersection-over-union score leaves a lot of room for local variation in the smoothness and shape of a 3-D model. So Wu and his colleagues also conducted a qualitative study of the models' fidelity to the source images. Of the study's participants, 74 percent preferred the new system's reconstructions to those of its predecessors.

All that fall

In another of Wu and Tenenbaum's papers, on which they're joined again by Freeman and by researchers at MIT, Cambridge University, and ShanghaiTech University, they train a system to analyze audio recordings of an object being dropped, to infer properties such as the object's shape, its composition, and the height from which it fell. Again, the system is trained to produce an abstract representation of the object, which, in turn, it uses to synthesize the sound the object would make when dropped from a particular height. The system's performance is judged on the similarity between the synthesized sound and the source sound.

Finally, in their fourth paper, Wu, Tenenbaum, Freeman, and colleagues at DeepMind and Oxford University describe a system that begins to model humans' intuitive understanding of the physical forces acting on objects in the world. This paper picks up where the previous papers leave off: It assumes that the system has already deduced objects' 3-D shapes.

Those shapes are simple: balls and cubes. The researchers trained their system to perform two tasks. The first is to estimate the velocities of balls traveling on a billiard table and, on that basis, to predict how they will behave after a collision. The second is to analyze a static image of stacked

cubes and determine whether they will fall and, if so, where the cubes will land.

Wu developed a representational language he calls scene XML that can quantitatively characterize the relative positions of objects in a visual scene. The system first learns to describe input data in that language. It then feeds that description to something called a physics engine, which models the physical forces acting on the represented objects. Physics engines are a staple of both computer animation, where they generate the movement of clothing, falling objects, and the like, and of scientific computing, where they're used for large-scale physical simulations.

After the physics engine has predicted the motions of the balls and boxes, that information is fed to a graphics engine, whose output is, again, compared with the source images. As with the work on visual discrimination, the researchers train their system on synthetic data before refining it with real data.

In tests, the researchers' system again outperformed its predecessors. In fact, in some of the tests involving billiard balls, it frequently outperformed human observers as well.

More information: Learning to See Physics via Visual De-animation. jiajunwu.com/papers/vda_nips.pdf

MarrNet: 3D Shape Reconstruction via 2.5D Sketches. jiajunwu.com/papers/marrnet_nips.pdf

Self-Supervised Intrinsic Image Decomposition. jiajunwu.com/papers/rin_nips.pdf

Shape and Material from Sound. jiajunwu.com/papers/fastsound_nips.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

APA citation: Computer systems predict objects' responses to physical forces (2017, December 14)

retrieved 21 January 2022 from https://techxplore.com/news/2017-12-responses-physical.html