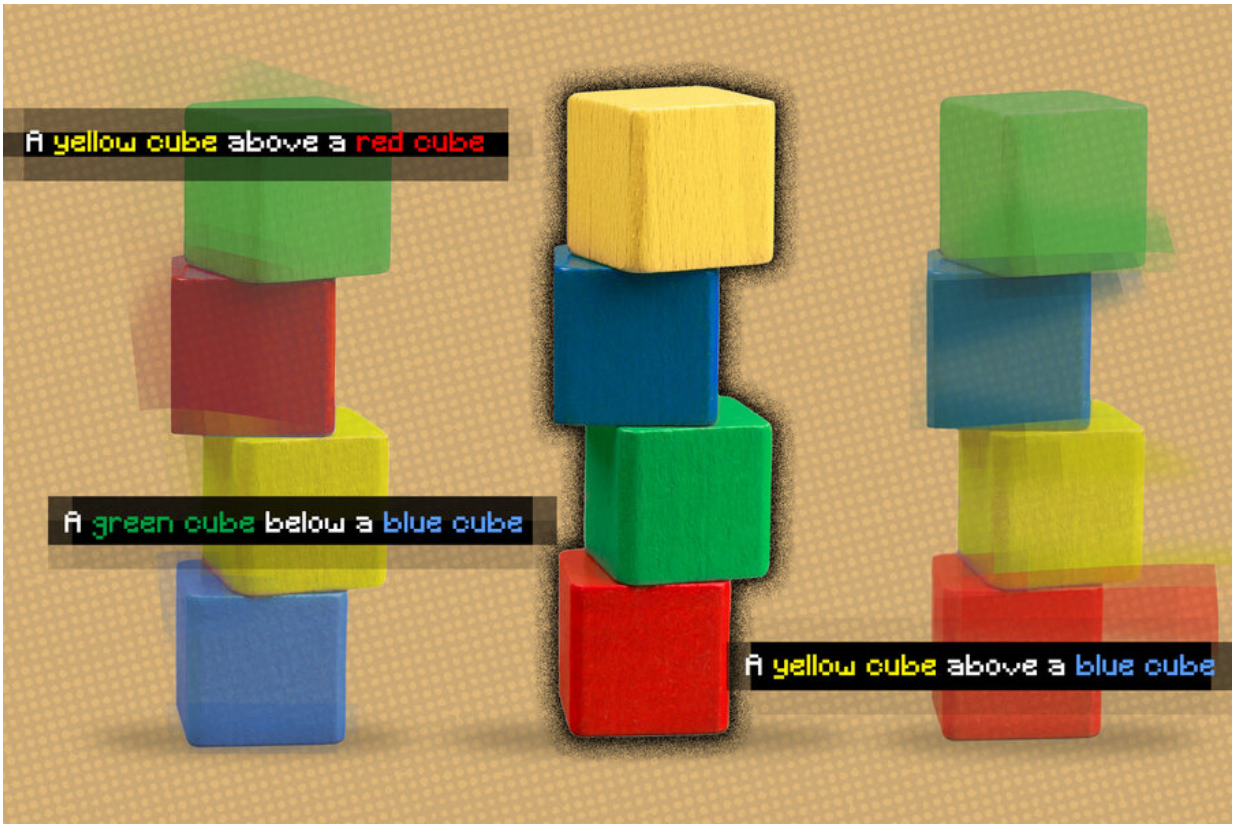
Machine-learning model could enable robots to understand interactions in the way humans do

November 29 2021, by Adam Zewe



MIT researchers have developed a machine learning model that understands the underlying relationships between objects in a scene and can generate accurate images of scenes from text descriptions. Credit: Jose-Luis Olivares, MIT, and iStockphoto

When humans look at a scene, they see objects and the relationships between them. On top of your desk, there might be a laptop that is sitting to the left of a phone, which is in front of a computer monitor.

Many deep learning models struggle to see the world this way because they don't understand the entangled relationships between individual objects. Without knowledge of these relationships, a robot designed to help someone in a kitchen would have difficulty following a command like "pick up the spatula that is to the left of the stove and place it on top of the cutting board."

In an effort to solve this problem, MIT researchers have developed a model that understands the underlying relationships between objects in a scene. Their model represents individual relationships one at a time, then combines these representations to describe the overall scene. This enables the model to generate more accurate images from text descriptions, even when the scene includes several objects that are arranged in different relationships with one another.

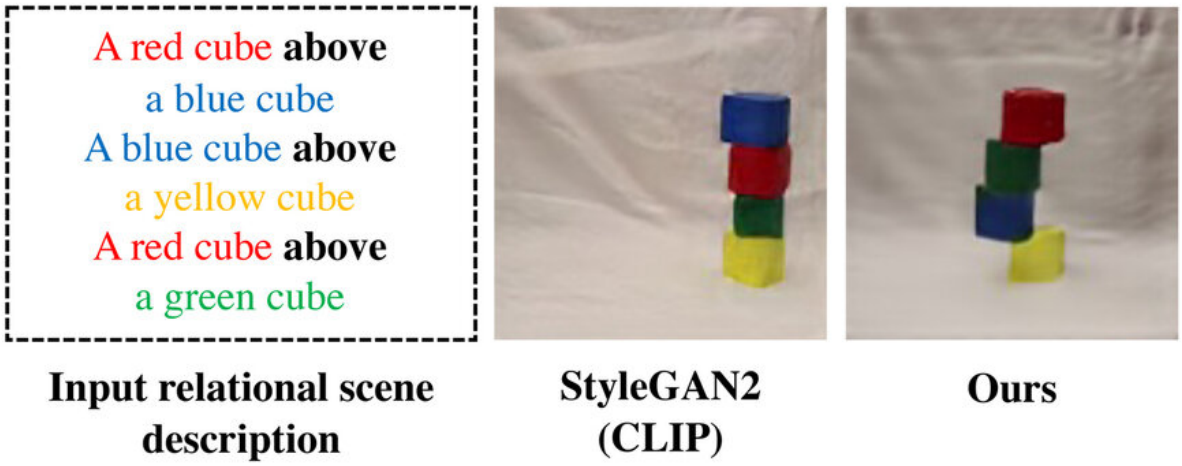
This work could be applied in situations where industrial robots must perform intricate, multistep manipulation tasks, like stacking items in a warehouse or assembling appliances. It also moves the field one step closer to enabling machines that can learn from and interact with their environments more like humans do.

"When I look at a table, I can't say that there is an object at XYZ location. Our minds don't work like that. In our minds, when we understand a scene, we really understand it based on the relationships between the objects. We think that by building a system that can understand the relationships between objects, we could use that system to more effectively manipulate and change our environments," says Yilun Du, a Ph.D. student in the Computer Science and Artificial Intelligence Laboratory (CSAIL) and co-lead author of the paper.

Du wrote the paper with co-lead authors Shuang Li, a CSAIL Ph.D. student, and Nan Liu, a graduate student at the University of Illinois at Urbana-Champaign; as well as Joshua B. Tenenbaum, the Paul E. Newton Career Development Professor of Cognitive Science and Computation in the Department of Brain and Cognitive Sciences and a member of CSAIL; and senior author Antonio Torralba, the Delta Electronics Professor of Electrical Engineering and Computer Science and a member of CSAIL. The research will be presented at the Conference on Neural Information Processing Systems in December.

One relationship at a time

The framework the researchers developed can generate an image of a scene based on a text description of objects and their relationships, like "A wood table to the left of a blue stool. A red couch to the right of a blue stool."



The framework the researchers developed can generate an image of a scene based on a text description of objects and their relationships, In this figure, researchers' final image is on the right and correctly follows the text description. Credit: Massachusetts Institute of Technology

Their system would break these sentences down into two smaller pieces that describe each individual relationship ("a wood table to the left of a blue stool" and "a red couch to the right of a blue stool"), and then model each part separately. Those pieces are then combined through an optimization process that generates an image of the scene.

The researchers used a machine-learning technique called energy-based models to represent the individual object relationships in a scene description. This technique enables them to use one energy-based model to encode each relational description, and then compose them together in a way that infers all objects and relationships.

By breaking the sentences down into shorter pieces for each relationship, the system can recombine them in a variety of ways, so it is better able to adapt to scene descriptions it hasn't seen before, Li explains.

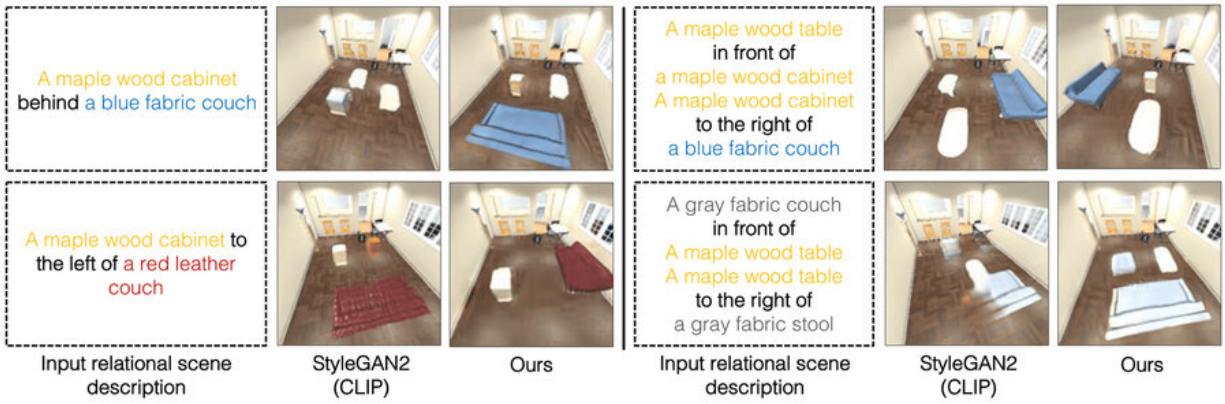
"Other systems would take all the relations holistically and generate the image one-shot from the description. However, such approaches fail when we have out-of-distribution descriptions, such as descriptions with more relations, since these model can't really adapt one shot to generate images containing more relationships. However, as we are composing these separate, smaller models together, we can model a larger number of relationships and adapt to novel combinations," Du says.

The system also works in reverse—given an image, it can find text descriptions that match the relationships between objects in the scene. In addition, their model can be used to edit an image by rearranging the objects in the scene so they match a new description.

Understanding complex scenes

The researchers compared their model to other deep learning methods that were given text descriptions and tasked with generating images that displayed the corresponding objects and their relationships. In each instance, their model outperformed the baselines.

They also asked humans to evaluate whether the generated images matched the original scene description. In the most complex examples, where descriptions contained three relationships, 91 percent of participants concluded that the new model performed better.



In this figure, the researcher's final images are labeled "ours." Credit: Massachusetts Institute of Technology

"One interesting thing we found is that for our model, we can increase our sentence from having one relation description to having two, or three, or even four descriptions, and our approach continues to be able to generate images that are correctly described by those descriptions, while other methods fail," Du says.

The researchers also showed the model images of scenes it hadn't seen

before, as well as several different text descriptions of each image, and it was able to successfully identify the description that best matched the object relationships in the image.

And when the researchers gave the system two relational scene descriptions that described the same image but in different ways, the model was able to understand that the descriptions were equivalent.

The researchers were impressed by the robustness of their model, especially when working with descriptions it hadn't encountered before.

"This is very promising because that is closer to how humans work. Humans may only see several examples, but we can extract useful information from just those few examples and combine them together to create infinite combinations. And our model has such a property that allows it to learn from fewer data but generalize to more complex scenes or image generations," Li says.

While these early results are encouraging, the researchers would like to see how their model performs on real-world images that are more complex, with noisy backgrounds and objects that are blocking one another.

They are also interested in eventually incorporating their model into robotics systems, enabling a robot to infer object relationships from videos and then apply this knowledge to manipulate objects in the world.

"Developing visual representations that can deal with the compositional nature of the world around us is one of the key open problems in computer vision. This paper makes significant progress on this problem by proposing an energy-based model that explicitly models multiple relations among the objects depicted in the image. The results are really impressive," says Josef Sivic, a distinguished researcher at the Czech

Institute of Informatics, Robotics, and Cybernetics at Czech Technical University, who was not involved with this research.

More information: Learning to Compose Visual Relations. composevisualrelations.github.io/

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

Citation: Machine-learning model could enable robots to understand interactions in the way humans do (2021, November 29) retrieved 29 March 2023 from https://techxplore.com/news/2021-11-machine-learning-enable-robots-interactions-humans.html