Machines that see the world more like humans do

December 8 2021, by Adam Zewe

This image shows how 3DP3 (bottom row) infers more accurate pose estimates of objects from input images (top row) than deep learning systems (middle row). Credit: Massachusetts Institute of Technology

Computer vision systems sometimes make inferences about a scene that

fly in the face of common sense. For example, if a robot were processing a scene of a dinner table, it might completely ignore a bowl that is visible to any human observer, estimate that a plate is floating above the table, or misperceive a fork to be penetrating a bowl rather than leaning against it.

Move that computer vision system to a self-driving car and the stakes become much higher —for example, such systems have failed to detect emergency vehicles and pedestrians crossing the street.

To overcome these errors, MIT researchers have developed a framework that helps machines see the world more like humans do. Their new artificial intelligence system for analyzing scenes learns to perceive real-world objects from just a few images, and perceives scenes in terms of these learned objects.

The researchers built the framework using probabilistic programming, an AI approach that enables the system to cross-check detected objects against input data, to see if the images recorded from a camera are a likely match to any candidate scene. Probabilistic inference allows the system to infer whether mismatches are likely due to noise or to errors in the scene interpretation that need to be corrected by further processing.

This common-sense safeguard allows the system to detect and correct many errors that plague the "deep-learning" approaches that have also been used for computer vision. Probabilistic programming also makes it possible to infer probable contact relationships between objects in the scene, and use common-sense reasoning about these contacts to infer more accurate positions for objects.

"If you don't know about the contact relationships, then you could say that an object is floating above the table—that would be a valid explanation. As humans, it is obvious to us that this is physically

unrealistic and the object resting on top of the table is a more likely pose of the object. Because our reasoning system is aware of this sort of knowledge, it can infer more accurate poses. That is a key insight of this work," says lead author Nishad Gothoskar, an electrical engineering and computer science (EECS) Ph.D. student with the Probabilistic Computing Project.

In addition to improving the safety of self-driving cars, this work could enhance the performance of computer perception systems that must interpret complicated arrangements of objects, like a robot tasked with cleaning a cluttered kitchen.

Gothoskar's co-authors include recent EECS Ph.D. graduate Marco Cusumano-Towner; research engineer Ben Zinberg; visiting student Matin Ghavamizadeh; Falk Pollok, a software engineer in the MIT-IBM Watson AI Lab; recent EECS master's graduate Austin Garrett; Dan Gutfreund, a principal investigator in the MIT-IBM Watson AI Lab; Joshua B. Tenenbaum, the Paul E. Newton Career Development Professor of Cognitive Science and Computation in the Department of Brain and Cognitive Sciences (BCS) and a member of the Computer Science and Artificial Intelligence Laboratory; and senior author Vikash K. Mansinghka, principal research scientist and leader of the Probabilistic Computing Project in BCS. The research is being presented at the Conference on Neural Information Processing Systems in December.

A blast from the past

To develop the system, called "3D Scene Perception via Probabilistic Programming (3DP3)," the researchers drew on a concept from the early days of AI research, which is that computer vision can be thought of as the "inverse" of computer graphics.

Computer graphics focuses on generating images based on the representation of a scene; computer vision can be seen as the inverse of this process. Gothoskar and his collaborators made this technique more learnable and scalable by incorporating it into a framework built using probabilistic programming.

"Probabilistic programming allows us to write down our knowledge about some aspects of the world in a way a computer can interpret, but at the same time, it allows us to express what we don't know, the uncertainty. So, the system is able to automatically learn from data and also automatically detect when the rules don't hold," Cusumano-Towner explains.

In this case, the model is encoded with prior knowledge about 3D scenes. For instance, 3DP3 "knows" that scenes are composed of different objects, and that these objects often lay flat on top of each other—but they may not always be in such simple relationships. This enables the model to reason about a scene with more common sense.

Learning shapes and scenes

To analyze an image of a scene, 3DP3 first learns about the objects in that scene. After being shown only five images of an object, each taken from a different angle, 3DP3 learns the object's shape and estimates the volume it would occupy in space.

"If I show you an object from five different perspectives, you can build a pretty good representation of that object. You'd understand its color, its shape, and you'd be able to recognize that object in many different scenes," Gothoskar says.

Mansinghka adds, "This is way less data than deep-learning approaches. For example, the Dense Fusion neural object detection system requires

thousands of training examples for each object type. In contrast, 3DP3 only requires a few images per object, and reports uncertainty about the parts of each objects' shape that it doesn't know."

The 3DP3 system generates a graph to represent the scene, where each object is a node and the lines that connect the nodes indicate which objects are in contact with one another. This enables 3DP3 to produce a more accurate estimation of how the objects are arranged. (Deep-learning approaches rely on depth images to estimate object poses, but these methods don't produce a graph structure of contact relationships, so their estimations are less accurate.)

Outperforming baseline models

The researchers compared 3DP3 with several deep-learning systems, all tasked with estimating the poses of 3D objects in a scene.

In nearly all instances, 3DP3 generated more accurate poses than other models and performed far better when some objects were partially obstructing others. And 3DP3 only needed to see five images of each object, while each of the baseline models it outperformed needed thousands of images for training.

When used in conjunction with another model, 3DP3 was able to improve its accuracy. For instance, a deep-learning model might predict that a bowl is floating slightly above a table, but because 3DP3 has knowledge of the contact relationships and can see that this is an unlikely configuration, it is able to make a correction by aligning the bowl with the table.

"I found it surprising to see how large the errors from deep learning could sometimes be—producing scene representations where objects really didn't match with what people would perceive. I also found it

surprising that only a little bit of model-based inference in our causal probabilistic program was enough to detect and fix these errors. Of course, there is still a long way to go to make it fast and robust enough for challenging real-time vision systems—but for the first time, we're seeing probabilistic programming and structured causal models improving robustness over deep learning on hard 3D vision benchmarks," Mansinghka says.

In the future, the researchers would like to push the system further so it can learn about an object from a single image, or a single frame in a movie, and then be able to detect that object robustly in different scenes. They would also like to explore the use of 3DP3 to gather training data for a neural network. It is often difficult for humans to manually label images with 3D geometry, so 3DP3 could be used to generate more complex image labels.

The 3DP3 system "combines low-fidelity graphics modeling with common-sense reasoning to correct large scene interpretation errors made by deep learning neural nets. This type of approach could have broad applicability as it addresses important failure modes of deep learning. The MIT researchers' accomplishment also shows how probabilistic programming technology previously developed under DARPA's Probabilistic Programming for Advancing Machine Learning (PPAML) program can be applied to solve central problems of common-sense AI under DARPA's current Machine Common Sense (MCS) program," says Matt Turek, DARPA Program Manager for the Machine Common Sense Program, who was not involved in this research, though the program partially funded the study.

More information: Nishad Gothoskar et al, 3DP3: 3D Scene Perception via Probabilistic Programming. arXiv:2111.00312v1 [cs.CV], arxiv.org/abs/2111.00312