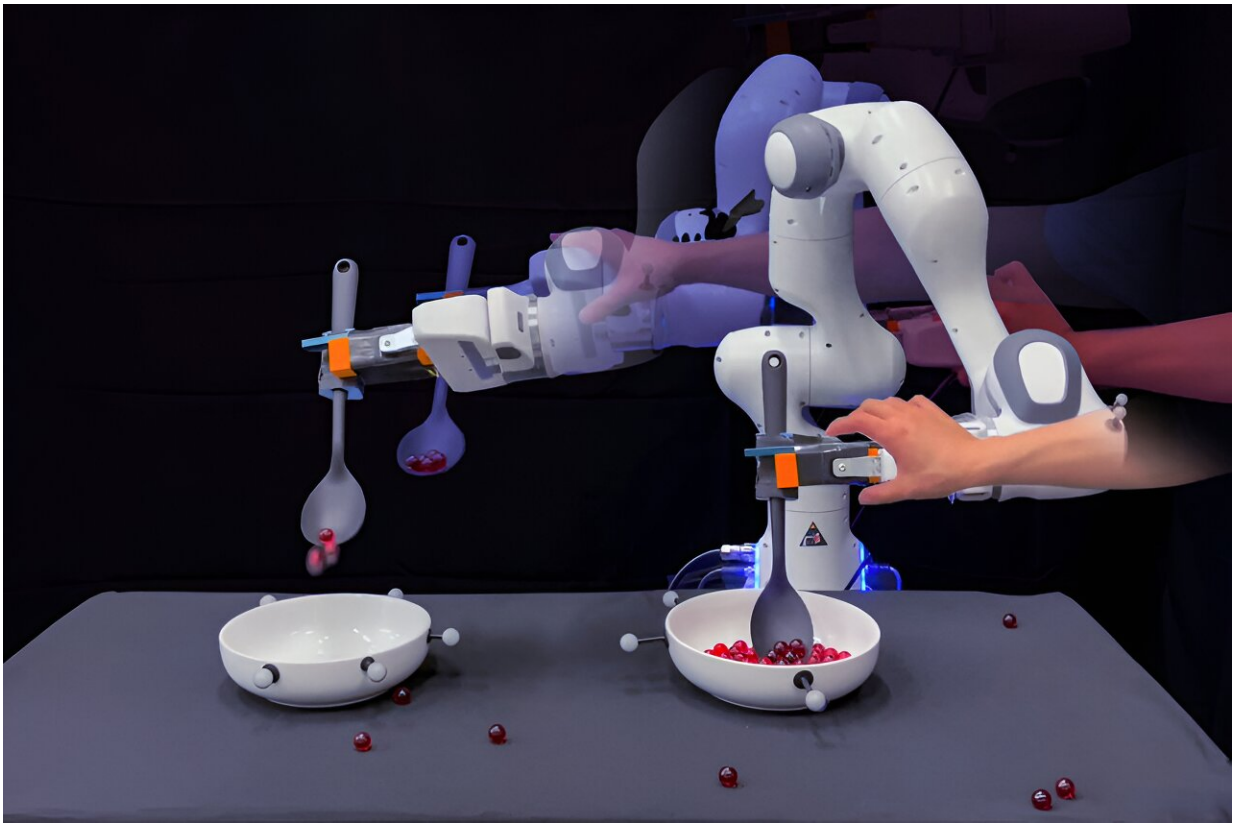
Engineering household robots to have a little common sense

March 25 2024, by Jennifer Chu



In this collaged image, a robotic hand tries to scoop up red marbles and put them into another bowl while a researcher's hand frequently disrupts it. The robot eventually succeeds. Credit: Jose-Luis Olivares, MIT. Stills courtesy of the researchers

From wiping up spills to serving up food, robots are being taught to carry

out increasingly complicated household tasks. Many such home-bot trainees are learning through imitation; they are programmed to copy the motions that a human physically guides them through.

It turns out that robots are excellent mimics. But unless engineers also program them to adjust to every possible bump and nudge, robots don't necessarily know how to handle these situations, short of starting their task from the top.

Now MIT engineers are aiming to give robots a bit of common sense when faced with situations that push them off their trained path. They've developed a method that connects robot motion data with the "common sense knowledge" of large language models, or LLMs.

Their approach enables a robot to logically parse many given household task into subtasks, and to physically adjust to disruptions within a subtask so that the robot can move on without having to go back and start a task from scratch—and without engineers having to explicitly program fixes for every possible failure along the way.

"Imitation learning is a mainstream approach enabling household robots. But if a robot is blindly mimicking a human's motion trajectories, tiny errors can accumulate and eventually derail the rest of the execution," says Yanwei Wang, a graduate student in MIT's Department of Electrical Engineering and Computer Science (EECS). "With our method, a robot can self-correct execution errors and improve overall task success."

Wang and his colleagues detail their new approach in a study they will present at the International Conference on Learning Representations (ICLR 2024) in May. The study's co-authors include EECS graduate students Tsun-Hsuan Wang and Jiayuan Mao, Michael Hagenow, a postdoc in MIT's Department of Aeronautics and Astronautics (AeroAstro), and Julie Shah, the H.N. Slater Professor in Aeronautics

and Astronautics at MIT.

Language task

The researchers illustrate their new approach with a simple chore: scooping marbles from one bowl and pouring them into another. To accomplish this task, engineers would typically move a robot through the motions of scooping and pouring—all in one fluid trajectory. They might do this multiple times, to give the robot a number of human demonstrations to mimic.

"But the human demonstration is one long, continuous trajectory," Wang says.

The team realized that, while a human might demonstrate a single task in one go, that task depends on a sequence of subtasks, or trajectories. For instance, the robot has to first reach into a bowl before it can scoop, and it must scoop up marbles before moving to the empty bowl, and so forth.

If a robot is pushed or nudged to make a mistake during any of these subtasks, its only recourse is to stop and start from the beginning, unless engineers were to explicitly label each subtask and program or collect new demonstrations for the robot to recover from the said failure, to enable a robot to self-correct in the moment.

"That level of planning is very tedious," Wang says.

Instead, he and his colleagues found some of this work could be done automatically by LLMs. These deep learning models process immense libraries of text, which they use to establish connections between words, sentences, and paragraphs. Through these connections, an LLM can then generate new sentences based on what it has learned about the kind of word that is likely to follow the last.

For their part, the researchers found that in addition to sentences and paragraphs, an LLM can be prompted to produce a logical list of subtasks that would be involved in a given task. For instance, if queried to list the actions involved in scooping marbles from one bowl into another, an LLM might produce a sequence of verbs such as "reach," "scoop," "transport," and "pour."

"LLMs have a way to tell you how to do each step of a task, in natural language. A human's continuous demonstration is the embodiment of those steps, in physical space," Wang says. "And we wanted to connect the two, so that a robot would automatically know what stage it is in a task, and be able to replan and recover on its own."

Mapping marbles

For their new approach, the team developed an algorithm to automatically connect an LLM's natural language label for a particular subtask with a robot's position in physical space or an image that encodes the robot state. Mapping a robot's physical coordinates, or an image of the robot state, to a natural language label is known as "grounding." The team's new algorithm is designed to learn a grounding "classifier," meaning that it learns to automatically identify what semantic subtask a robot is in—for example, "reach" versus "scoop"—given its physical coordinates or an image view.

"The grounding classifier facilitates this dialogue between what the robot is doing in the physical space and what the LLM knows about the subtasks, and the constraints you have to pay attention to within each subtask," Wang explains.

The team demonstrated the approach in experiments with a robotic arm that they trained on a marble-scooping task. Experimenters trained the robot by physically guiding it through the task of first reaching into a

bowl, scooping up marbles, transporting them over an empty bowl, and pouring them in.

After a few demonstrations, the team then used a pretrained LLM and asked the model to list the steps involved in scooping marbles from one bowl to another. The researchers then used their new algorithm to connect the LLM's defined subtasks with the robot's motion trajectory data. The algorithm automatically learned to map the robot's physical coordinates in the trajectories and the corresponding image view to a given subtask.

The team then let the robot carry out the scooping task on its own, using the newly learned grounding classifiers. As the robot moved through the steps of the task, the experimenters pushed and nudged the bot off its path, and knocked marbles off its spoon at various points.

Rather than stop and start from the beginning again, or continue blindly with no marbles on its spoon, the bot was able to self-correct, and completed each subtask before moving on to the next. (For instance, it would make sure that it successfully scooped marbles before transporting them to the empty bowl.)

"With our method, when the robot is making mistakes, we don't need to ask humans to program or give extra demonstrations of how to recover from failures," Wang says. "That's super exciting because there's a huge effort now toward training household robots with data collected on teleoperation systems. Our algorithm can now convert that training data into robust robot behavior that can do complex tasks, despite external perturbations."

More information: Paper submission: Grounding Language Plans in

Demonstrations Through Counter-Factual Perturbations

Yanwei Wang et al, Grounding Language Plans in Demonstrations Through Counterfactual Perturbations, *arXiv* (2024). DOI: 10.48550/arxiv.2403.17124 , arxiv.org/abs/2403.17124