Making sense of AI

Can you teach AI common sense?

Ben Dickson

@BenDee983

July 27, 2021 2:20 PM

Image Credit: Peach_iStock/Getty Images

Even before they speak their first words, human babies develop mental models about objects and people. This is one of the key capabilities that allows us humans to learn to live socially and cooperate (or compete) with each other. But for artificial intelligence, even the most basic behavioral reasoning tasks remain a challenge.

Advanced deep learning models can do complicated tasks such as detect people and objects in images, sometimes even better than humans. But they struggle to move beyond the visual features of images and make inferences about what other agents are doing or wish to accomplish.

To help fill this gap, scientists at IBM, the Massachusetts Institute of Technology, and Harvard University have developed a series of tests that will help evaluate the capacity of AI models to reason like children, by observing and making sense of the world

"Like human infants, it is critical for machine agents to develop an adequate capacity of understanding human minds, in order to successfully engage in social interactions," the AI researchers write in a new paper that introduces the dataset, called AGENT.

Presented at this year's International Conference on Machine Learning (ICML), AGENT provides an important benchmark for measuring the reasoning capabilities of AI systems.

Observing and predicting agent behavior

There's a large body of work on testing common sense and reasoning in AI systems. Many of them are focused on natural language understanding, including the famous Turing Test and Winograd schemas. In contrast, the AGENT project focuses on the kinds of reasoning capabilities humans learn before being able to speak.

"Our goal, following the literature in developmental psychology, is to create a benchmark for evaluating specific commonsense capabilities related to intuitive psychology which babies learn during the pre-lingual stage (in the first 18 months of their lives)," Dan Gutfreund, principal investigator at the MIT-IBM Watson AI Lab, told *TechTalks*.

As children, we learn to tell the difference between objects and agents by observing our environments. As we watch events unfold, we develop intuitive psychological skills, predict the goals of other people by observing their actions, and continue to correct and update our mental. We learn all this with little or no instructions.

The idea behind the AGENT (Action, Goal, Efficiency, coNstraint, uTility) test is to assess how well AI systems can mimic this basic skill, what they can develop psychological reasoning capabilities, and how well the representations they learn generalize to novel situations. The dataset comprises short sequences that show an agent navigating its way toward one of several objects. The sequences have been produced in ThreeDWorld, a virtual 3D environment designed for training AI agents.

The AGENT test takes place in two phases. First, the AI is presented with one or two sequences that depict the agent's behavior. These examples should familiarize the AI with the virtual agent's preferences. For example, an agent might always choose one type of object regardless of the obstacles that stand in its way, or it might choose the closest and most accessible object regardless of its type.

After the familiarization phase, the AI is shown a test sequence and it must determine whether the agent is acting in an expected or surprising manner.

The tests, 3,360 in total, span across four types of scenarios, starting with very simple behavior (the agent prefers one type of object regardless of the environment) to more complicated challenges (the agent manifests cost-reward estimation, weighing the difficulty of achieving a goal against the reward it will receive). The AI must also consider

there are no obstacles). And in some of the challenges, the scene is partially occluded to make it more difficult to reason about the environment.

Realistic scenarios in an artificial environment

The designers of the tests have included human inductive biases, which means the agents and environment are governed by rules that would be rational to humans (e.g., the cost of jumping or climbing an obstacle grows with its height). This decision helps make the challenges more realistic and easier to evaluate. The researchers also note that these kinds of biases are also important to help create AI systems that are better aligned and compatible with human behavior and can cooperate with human counterparts.

The AI researchers tested the challenges on human volunteers through Amazon Mechanical Turk. Their findings show that on average, humans can solve 91 percent of the challenges by observing the familiarization sequences and judging the test examples. This implies that humans use their prior knowledge about the world and human/animal behavior to make sense of how the agents make decision (e.g., all other things being equal, an agent will choose the object with higher reward).

The AI researchers intentionally limited the size of the dataset to prevent unintelligent shortcuts to solving the problems. Given a very large dataset, a machine learning model

about agent behavior. “Training from scratch on just our dataset will not work. Instead, we suggest that to pass the tests, it is necessary to acquire additional knowledge either via inductive biases in the architectures, or from training on additional data,” the researchers write.

The researchers, however, have implemented some shortcuts in the tests. The AGENT dataset includes depth maps, segmentation maps, and bounding boxes of objects and obstacles for every frame of the scene. The scenes are also extremely simple in visual details and are composed of eight distinct colors. All of this makes it easier for AI systems to process the information in the scene and focus on the reasoning part of the challenge.

Does current AI solve AGENT challenges?

The researchers tested the AGENT challenge on two baseline AI models. The first one, Bayesian Inverse Planning and Core Knowledge (BIPaCK), is a generative model that integrates physics simulation and planning.

Above: The BIPaCK model uses planner and physics engines to predict the trajectory of the agent

This model uses the full ground-truth information provided by the dataset and feeds it into its physics and planning engine to predict the trajectory of the agent. The researchers’ experiments show that BIPaCK is able to perform on par or even better than

However, in the real world, AI systems don't have access to precisely annotated ground truth information and must perform the complicated task of detecting objects against different backgrounds and lighting conditions, a problem that humans and animals solve easily but remains a challenge for computer vision systems.

In their paper, the researchers acknowledge that the BIPaCK "requires an accurate reconstruction of the 3D state and a built-in model of the physical dynamics, which will not necessarily be available in real world scenes."

The second model the researchers tested, codenamed ToMnet-G, is an extended version of the Theory of Mind Neural Network (ToMnet), proposed by scientists at DeepMind in 2018. ToMnet-G uses graph neural networks to encode the state of the scenes, including the objects, obstacles, and the agent's location. It then feeds those encodings into long short-term memory networks (LSTM) to track the agent's trajectory across the sequence of frames. The model uses the representations it extracts from the familiarization videos to predict the agent's behavior in the test videos and rate them as expected or surprising.

Above: The ToMnet-G model uses graph neural networks and LSTMs to embed scene representations and predict agent behavior

The advantage of ToMnet-G is that it does not require the pre-engineered physics and commonsense knowledge of BIPaCK. It learns everything from the videos and previous training on other datasets. On the other hand, ToMnet-G often learns the wrong representations and can't generalize its behavior to new scenarios or when it has limited familiarity information.

“Without many built-in priors, ToMnet-G demonstrates promising results when trained and tested on similar scenarios, but it still lacks a strong generalization capacity both within scenarios and across them,” the researchers observe in their paper.

The contrast between the two models highlights the challenges of the simplest tasks that humans learn without any instructions.

“We have to remember that our benchmark, by design, depicts very simple synthetic scenarios addressing each time one specific aspect of common sense,” Gutfreund said. “In the real world, humans are able to very quickly parse complex scenes where simultaneously many aspects of common sense related to physics, psychology, language and more are at play. AI models are still far from being able to do anything close to that.”




Common sense and the future of AI

"We believe that the path from narrow to broad AI has to include models that have common sense," Gutfreund said. "Common sense capabilities are important building blocks in understanding and interacting in the world and can facilitate the acquisition of new capabilities."

Many scientists believe that common sense and reasoning can solve many of the problems current AI systems face, such as their need for extensive volumes of training data, their struggle with causality, and their fragility in dealing with novel situations. Common sense and reasoning are important areas of research for the AI community, and they have become the focus of some of the brightest minds in the field, including the pioneers of deep learning.

Solving AGENT can be a small but important step toward creating AI agents that behave robustly in the unpredictable world of humans.

"It will be difficult to convince people to trust autonomous agents which do not behave in a common sensical way," Gutfreund said. "Consider, for example, a robot for assisting the elderly. If that robot will not follow the commonsense principal that agents pursue their goals efficiently and will move in zig zag rather than in a straight line when asked to fetch milk from the fridge, it will not be very practical nor trustworthy."

AGENT is part of the Machine Common Sense (MCS) program of the Defense Advanced Research Projects Agency (DARPA). MCS follows two broad goals. The first is to create machines that can learn like children to reason about objects, agents, and space. AGENT falls into this category. The second goal is to develop systems that can learn by reading structured and unstructured knowledge from the web, as a human researcher would do. This is different from current approaches to natural language understanding, which focus only on capturing statistical correlations between words and word sequences in very large corpora of text.

"We are now working on using AGENT as a testing environment for babies. Together with the rest of the DARPA MCS program performers we are planning to explore more complex scenarios of common sense related to multiple agents (e.g., helping or hindering each

other core domains of knowledge related to intuitive physics and spatial understanding," Gutfreund said.

Ben Dickson is a software engineer and the founder of TechTalks, a blog that explores the ways technology is solving and creating problems.