Is this a deer I see? Socially aware AI adapts by asking questions of humans

November 10 2022, by Andrew Myers

Scholars built an agent that learns by asking humans questions and adapts them based on socially aware observations. If "What type of animal is that?" elicits no response, the algorithm might ask instead, "Is that a deer I see?" Credit: DALL-E

As good as they've become, artificial intelligence agents are still largely only as good as the data upon which they were trained. They don't know

what they don't know. In the real world, people faced with unfamiliar situations and surroundings adapt by watching what others around them are doing and by asking questions. When in Rome, as they say. Experts in educational psychology call this "socially situated learning."

Until now, AI agents have lacked this ability to learn on the fly, but researchers at Stanford University recently announced that they have developed artificially intelligent agents with the ability to seek out new knowledge by asking people questions.

"We built an agent that looks at photos and learns to ask natural language questions about them to expand its knowledge beyond the datasets it was originally trained on," says Ranjay Krishna, first author of a recent study appearing in the journal *Proceedings of the National Academy of Sciences*. Krishna earned his doctorate at Stanford and is now on the faculty at the University of Washington.

Uncanny awareness

The new approach combines aspects of computer vision and human cognitive and behavioral sciences to take machine learning in a new direction. They call it "socially situated artificial intelligence."

The kicker in this research is that when people are unwilling or disinterested in responding to AI's questions, which can often seem simplistic or mundane, the AI adapts.

For instance, when analyzing a photo of a person and an unfamiliar four-legged animal, the algorithm might first ask, "What type of animal is that?" That might beget ironic or sarcastic answers ("That's a human.") Or, worse, it might get no answer at all. Instead, the algorithm might ask, "Is that a dog I see?" Posing the question in this way is more likely to engender a truthful answer: "No, that's a deer."

"Much as we'd like to think that people are earnest respondents, willing to answer any question the AI might pose, often they are not," Krishna says. "Our agent senses this and changes its questions based on its socially aware observations of which questions people will and won't answer."

The new agent achieves several goals at once. First, of course, it learns new visual concepts, which is the main goal. But, second, it also learns to read social norms. Additionally, Krishna notes, there is a cumulative effect. After asking questions and learning new information, the AI retrains itself. The next time, it asks different questions because it has learned more things about the world.

"There are only so many ways you can describe a table. A person might understandably not want to answer questions seen as disingenuous, nonsensical, or just plain boring," Krishna says. "But the agent gets around those challenges with clever questions that become more sophisticated as the agent learns."

Testing the hypothesis

To test their approach, the research team, including Krishna's doctoral advisors Stanford HAI Co-Director Fei-Fei Li and Michael Bernstein, professors in Stanford's Department of Computer Science, engaged in an eight-month experiment where their algorithm viewed images posted on a photography-based social media platform and asked questions of some 236,000 people, many of whom were the photographers themselves.

Over the course of the experiment, the new algorithm more than doubled its ability to recognize new visual information in images posted by its human correspondents.

The potential of social AI

Socially situated AI, the researchers believe, can overcome limitations on how AI learns and push intelligence gathering in new directions. The researchers think the approach creates opportunities for AI agents able to recognize their own anti-social behaviors and adapt questions on the fly to avoid that all-too-human quality: boredom.

"The agent has an iterative learning process where every once in a while, it uses all the new content that it has seen and retrains itself, so that the next time it would ask different questions based on the things it has learned about the world," Krishna says. "As it learns to ask better questions, the human respondents stay engaged."

Most important, it allows the AI to learn new information beyond the datasets upon which it was originally trained. Current methods of training are not unlike locking the agent in a room with a stack of books, as the authors note in their paper. Once those pages are learned, all future decisions must be made only using information gleaned from those books and nothing else.

Further complicating matters, whatever books the AI is given to train on must first be annotated by people—a process known in artificial intelligence as "labeling." That is, human annotators must tell the AI what it is seeing before the AI can learn to see. Unfortunately, getting annotators to label routine content is a challenge. And, without that labeling, the AI cannot learn.

Instead of asking annotators to label data, agents that can learn socially by asking questions about their situations are more likely to garner helpful responses from people.

The new agent effectively achieves labeling by asking questions and,

when it senses reluctance on the part of human correspondents, it learns to ask questions in new ways to get earnest, truthful answers.

There are certain potential risks as well. Pointing to Microsoft's Tay, a similar agent deployed on Twitter that soon began posting anti-social tweets learned from its interactions with people, this socially situated AI does not suffer from the same issues. Users do not initiate interactions and, therefore, cannot coordinate attacks on the agent. The agent decides whom to interact with and asks interesting questions to control what it learns.

The authors also conducted experiments to study how AIs should introduce themselves to people to garner helpful and avoid "troll" responses. Regardless, Krishna says there is still a lot of research to be done to account for the biases that AIs might learn from people and to mitigate the risks that those learned biases might result in.

As for now, the current version of the agent operates digitally on social media. The next step for Krishna is to transfer this approach to real-world situations in which a person might correct robots on the fly when it sees them making a mistake. He foresees a day when people might be able to teach robots, in their own homes, to accomplish new tasks that make their lives easier.

Other potential applications include health care, where robots might ask providers to clarify their medical procedures, technologies that modify their interfaces based on direct user feedback, and culturally aware agents that can learn from diverse communities to improve learning.

"I would be interested in moving into the physical world where people are interacting with robots to get them to solve new tasks. Or, if you see your AI making a mistake, people should be able to quickly provide feedback and correct it immediately," Krishna says of his next steps.

More information: Ranjay Krishna et al, Socially situated artificial intelligence enables learning from human interaction, *Proceedings of the National Academy of Sciences* (2022). DOI: 10.1073/pnas.2115730119

Provided by Stanford University