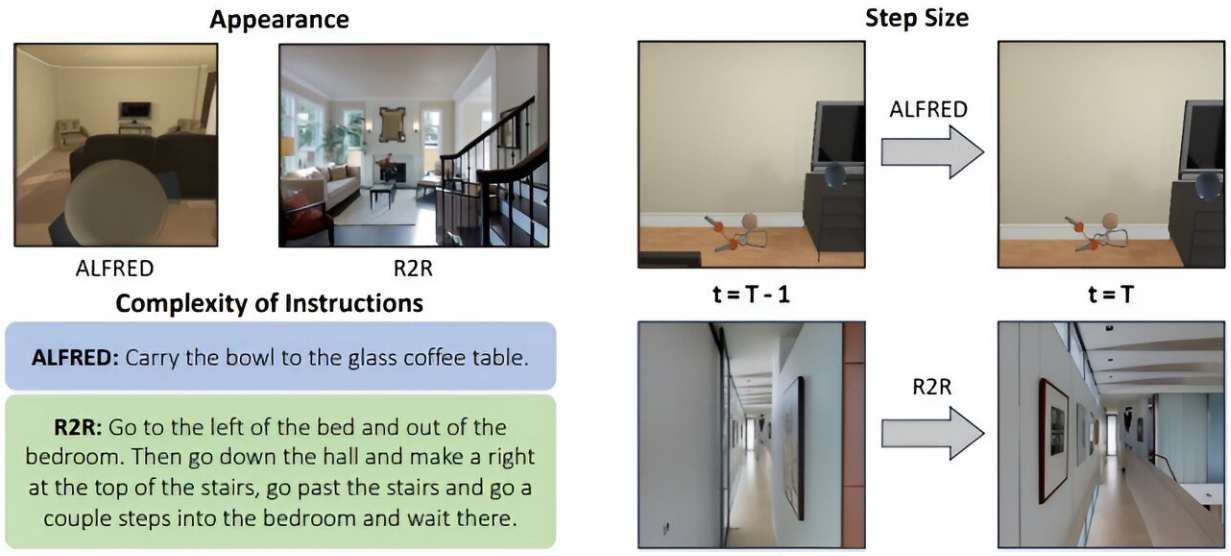
New method uses language-based inputs instead of costly visual data to help robots navigate

June 12 2024, by Adam Zewe



Task gap between ALFRED and R2R. We highlight notable distinctions between the navigation tasks in ALFRED and R2R, encompassing variations in appearance, step size, and instruction complexity. Credit: *arXiv* (2023). DOI: 10.48550/arxiv.2310.07889

Someday, you may want your home robot to carry a load of dirty clothes downstairs and deposit them in the washing machine in the far-left corner of the basement. The robot will need to combine your instructions with its visual observations to determine the steps it should take to

complete this task.

For an AI agent, this is easier said than done. Current approaches often utilize multiple hand-crafted machine-learning models to tackle different parts of the task, which require a great deal of human effort and expertise to build. These methods, which use visual representations to directly make navigation decisions, demand massive amounts of visual data for training, which are often hard to come by.

To overcome these challenges, researchers from MIT and the MIT-IBM Watson AI Lab devised a navigation method that converts visual representations into pieces of language, which are then fed into one large language model that achieves all parts of the multistep navigation task.

Rather than encoding visual features from images of a robot's surroundings as visual representations, which is computationally intensive, their method creates text captions that describe the robot's point-of-view. A large language model uses the captions to predict the actions a robot should take to fulfill a user's language-based instructions.

Because their method utilizes purely language-based representations, they can use a large language model to efficiently generate a huge amount of synthetic training data.

While this approach does not outperform techniques that use visual features, it performs well in situations that lack enough visual data for training. The researchers found that combining their language-based inputs with visual signals leads to better navigation performance.

"By purely using language as the perceptual representation, ours is a more straightforward approach. Since all the inputs can be encoded as language, we can generate a human-understandable trajectory," says Bowen Pan, an electrical engineering and computer science (EECS)

graduate student and lead author of a paper on this approach, which is published on the *arXiv* preprint server.

Solving a vision problem with language

Since large language models are the most powerful machine-learning models available, the researchers sought to incorporate them into the complex task known as vision-and-language navigation, Pan says.

But such models take text-based inputs and can't process visual data from a robot's camera. So, the team needed to find a way to use language instead.

Their technique utilizes a simple captioning model to obtain text descriptions of a robot's visual observations. These captions are combined with language-based instructions and fed into a large language model, which decides what navigation step the robot should take next.

The large language model outputs a caption of the scene the robot should see after completing that step. This is used to update the trajectory history so the robot can keep track of where it has been.

The model repeats these processes to generate a trajectory that guides the robot to its goal, one step at a time.

To streamline the process, the researchers designed templates so observation information is presented to the model in a standard form—as a series of choices the robot can make based on its surroundings.

For instance, a caption might say "to your 30-degree left is a door with a potted plant beside it, to your back is a small office with a desk and a computer," etc. The model chooses whether the robot should move

toward the door or the office.

"One of the biggest challenges was figuring out how to encode this kind of information into language in a proper way to make the agent understand what the task is and how they should respond," Pan says.

Advantages of language

When they tested this approach, while it could not outperform vision-based techniques, they found that it offered several advantages.

First, because text requires fewer computational resources to synthesize than complex image data, their method can be used to rapidly generate synthetic training data. In one test, they generated 10,000 synthetic trajectories based on 10 real-world, visual trajectories.

The technique can also bridge the gap that can prevent an agent trained with a simulated environment from performing well in the real world. This gap often occurs because computer-generated images can appear quite different from real-world scenes due to elements like lighting or color. But language that describes a synthetic versus a real image would be much harder to tell apart, Pan says.

Also, the representations their model uses are easier for a human to understand because they are written in natural language.

"If the agent fails to reach its goal, we can more easily determine where it failed and why it failed. Maybe the history information is not clear enough or the observation ignores some important details," Pan says.

In addition, their method could be applied more easily to varied tasks and environments because it uses only one type of input. As long as data can be encoded as language, they can use the same model without

making any modifications.

But one disadvantage is that their method naturally loses some information that would be captured by vision-based models, such as depth information.

However, the researchers were surprised to see that combining language-based representations with vision-based methods improves an agent's ability to navigate.

"Maybe this means that language can capture some higher-level information than cannot be captured with pure vision features," he says.

This is one area the researchers want to continue exploring. They also want to develop a navigation-oriented captioner that could boost the method's performance. In addition, they want to probe the ability of large language models to exhibit spatial awareness and see how this could aid language-based navigation.

More information: Bowen Pan et al, LangNav: Language as a Perceptual Representation for Navigation, *arXiv* (2023). DOI: 10.48550/arxiv.2310.07889

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

navigate (2024, June 12) retrieved 26 June 2024 from
https://techxplore.com/news/2024-06-method-language-based-visual-robots.html