TechXplore

Stress test method detects when object recognition models are using shortcuts

July 18 2023

From left to right, Sriram Yenamandra, Viraj Prabhu, and Prithvijit Chattopadhyay, discuss their LANCE method for detecting input changes that deep object recognition models are sensitive to. Credit: Kevin Beasley/College of Computing

A new "stress test" method created by a Georgia Tech researcher allows programmers to more easily determine if trained visual recognition models are sensitive to input changes or rely too heavily on context clues to perform their tasks.

Viraj Prabhu, a Ph.D. student in Georgia Tech's School of Interactive Computing, introduced the LANCE (Language-Guided Counterfactuals) method in a recent research paper published on the preprint server *arXiv* that shows how deep object recognition models are prone to taking shortcuts through context clues to produce images.

Ideally, models should understand exactly what they're prompted to search for, Prabhu said, but because of spurious correlation, they tend to use irrelevant information in images as they make predictions.

Prabhu used LANCE to stress test well-known models that have been trained on the image database ImageNet. Working with Assistant Professor Judy Hoffman and co-authors Sriram Yenamandra and Prithvijit Chattopadhyay, he discovered many instances in which the models were overly reliant on context in the images they produced.

In some examples, the models showed they were using weather in the background to classify images rather than recognizing the object of interest.

On another stress test, Prabhu challenged the models to classify images with seatbelts. All the test images contained seatbelts inside cars. When Prabhu generated new images by changing the parameters to "seatbelts on a bus," the performance and accuracy of the trained models dropped. This suggested the models thought seat belts were exclusive to cars.

"When a model is getting something right, is it getting it right because it really understands it, or is it picking up on some context clues and relying on them?" Prabhu said.

"There is no reason why it should be relying on what kind of vehicle it is to know whether there is a seatbelt, but models often do this. It's more generally known as model bias or a spurious correlation problem."

The models displayed the same flaws when Prabhu used LANCE to test images for dog sleds. The models almost exclusively associated dog sleds with Huskies, leading them to focus their searches on the breed most associated with sleds.

Prabhu said the prompts given to the models were generated by finetuning LLaMA, a large-language model created by Meta AI, while using training data automatically generated by Open AI's ChatGPT. For an image of someone riding a bike, he generated a caption using an automated captioning system. Then, he used the finetuned LLaMA to make a structured change to the caption, only changing a single concept at a time.

"It would change 'person riding a bicycle' to 'person carrying a bicycle,' and then we pass it to the generative model and use it to generate a new image while changing nothing else," he said. "Using a recently introduced targeted editing technique from Google Research based on prompt-to-prompt tuning, we can now change only the relationship between the person and bicycle. Then we get an image of a person carrying a bicycle, with everything else being the same. Now we can use this as a counterfactual test image."

That allows Prabhu to compare the model's new prediction to the original. If the prediction has changed, it's likely the model is relying on spurious correlations.

Prabhu said the LANCE method can be applied at scale for any new data set.

Spurious correlation has been a known weak link for deep learning models, but Prabhu said the benefit of LANCE is that it allows programmers to probe their models for those weaknesses before deployment.

Traditionally, these models are trained through goal-oriented methods in which the models receive points for displaying the correct image and lose points for getting them wrong. Prabhu said that's the most likely reason why the artificial intelligence in the models tries to find shortcuts, like using contextual clues, to achieve their goals.

The implications also expand beyond diagnosing object recognition models trained on ImageNet. LANCE can be applied to computer vision technology used in self-driving vehicles, which need to be as foolproof as possible before they're deployed on the road.

"In high-stakes applications like self-driving, people are using discriminative approaches—you have an object detection system that can detect cars and pedestrians and draw boxes around them," Prabhu said. "Using LANCE, we can probe these discriminative models using generative approaches and make them better. The hope is we can discover failures before they happen."

More information: Viraj Prabhu et al, LANCE: Stress-testing Visual Models by Generating Language-guided Counterfactual Images, *arXiv* (2023). DOI: 10.48550/arxiv.2305.19164

Provided by Georgia Institute of Technology

Citation: Stress test method detects when object recognition models are using shortcuts (2023, July 18) retrieved 2 May 2024 from

https://techxplore.com/news/2023-07-stress-method-recognition-shortcuts.html