More transparency and understanding into machine behaviors

March 25 2021, by Rachel Gordon

Credit: Pixabay/CC0 Public Domain

Explaining, interpreting, and understanding the human mind presents a unique set of challenges. Doing the same for the behaviors of machines, meanwhile, is a whole other story.

As artificial intelligence (AI) models are increasingly used in complex situations—approving or denying loans, helping doctors with medical diagnoses, assisting drivers on the road, or even taking complete control—humans still lack a holistic understanding of their capabilities and behaviors.

Existing research focuses mainly on the basics: How accurate is this model? Oftentimes, centering on the notion of simple accuracy can lead to dangerous oversights. What if the model makes mistakes with very high confidence? How would the model behave if it encountered something previously unseen, such as a self-driving car seeing a new type of traffic sign?

In the quest for better human-AI interaction, a team of researchers from MIT's Computer Science and Artificial Intelligence Laboratory (CSAIL) have created a new tool called Bayes-TrEx that allows developers and users to gain transparency into their AI model. Specifically, it does so by finding concrete examples that lead to a particular behavior. The method makes use of "Bayesian posterior inference," a widely-used mathematical framework to reason about model uncertainty.

In experiments, the researchers applied Bayes-TrEx to several image-based datasets, and found new insights that were previously overlooked by standard evaluations focusing solely on prediction accuracy.

"Such analyses are important to verify that the model is indeed functioning correctly in all cases," says MIT CSAIL Ph.D. student Yilun Zhou, co-lead researcher on Bayes-TrEx. "An especially alarming situation is when the model is making mistakes, but with very high confidence. Due to high user trust over the high reported confidence, these mistakes might fly under the radar for a long time and only get discovered after causing extensive damage."

For example, after a medical diagnosis system finishes learning on a set of X-ray images, a doctor can use Bayes-TrEx to find images that the model misclassified with very high confidence, to ensure that it doesn't miss any particular variant of a disease.

Bayes-TrEx can also help with understanding model behaviors in novel situations. Take autonomous driving systems, which often rely on camera images to take in traffic lights, bike lanes, and obstacles. These common occurrences can be easily recognized with high accuracy by the camera, but more complicated situations can provide literal and metaphorical roadblocks. A zippy Segway could potentially be interpreted as something as big as a car or as small as a bump on the road, leading to a tricky turn or costly collision. Bayes-TrEx could help address these novel situations ahead of time, and enable developers to correct any undesirable outcomes before potential tragedies occur.

In addition to images, the researchers are also tackling a less-static domain: robots. Their tool, called "RoCUS," inspired by Bayes-TrEx, uses additional adaptations to analyze robot-specific behaviors.

While still in a testing phase, experiments with RoCUS point to new discoveries that could be easily missed if the evaluation was focused solely on task completion. For example, a 2D navigation robot that used a deep learning approach preferred to navigate tightly around obstacles, due to how the training data was collected. Such a preference, however, could be risky if the robot's obstacle sensors are not fully accurate. For a robot arm reaching a target on a table, the asymmetry in the robot's kinematic structure showed larger implications on its ability to reach targets on the left versus the right.

"We want to make human-AI interaction safer by giving humans more insight into their AI collaborators," says MIT CSAIL Ph.D. student Serena Booth, co-lead author with Zhou. "Humans should be able to

understand how these agents make decisions, to predict how they will act in the world, and—most critically—to anticipate and circumvent failures."

More information: Bayes-TrEx: a Bayesian Sampling Approach to Model Transparency by Example. slbooth.com/BayesTrex_files/BayesTrex.html

Provided by Massachusetts Institute of Technology