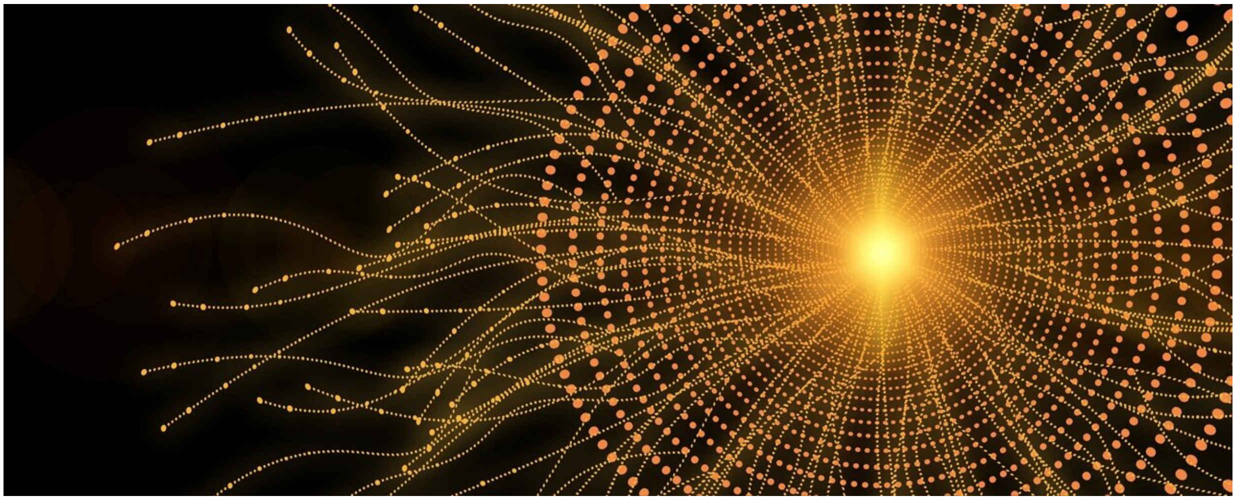
TechXplore

How to tell if artificial intelligence is working the way we want it to

July 22 2022, by Adam Zewe



Credit: Pixabay/CC0 Public Domain

About a decade ago, deep-learning models started achieving superhuman results on all sorts of tasks, from beating world-champion board game players to outperforming doctors at diagnosing breast cancer.

These powerful deep-learning models are usually based on artificial neural networks, which were first proposed in the 1940s and have become a popular type of machine learning. A computer learns to process data using layers of interconnected nodes, or neurons, that mimic the human brain.

As the field of machine learning has grown, artificial neural networks have grown along with it.

Deep-learning models are now often composed of millions or billions of interconnected nodes in many layers that are trained to perform detection or classification tasks using vast amounts of data. But because the models are so enormously complex, even the researchers who design them don't fully understand how they work. This makes it hard to know whether they are working correctly.

For instance, maybe a model designed to help physicians diagnose patients correctly predicted that a skin lesion was cancerous, but it did so by focusing on an unrelated mark that happens to frequently occur when there is cancerous tissue in a photo, rather than on the cancerous tissue itself. This is known as a spurious correlation. The model gets the prediction right, but it does so for the wrong reason. In a real clinical setting where the mark does not appear on cancer-positive images, it could result in missed diagnoses.

With so much uncertainty swirling around these so-called "black-box" models, how can one unravel what's going on inside the box?

This puzzle has led to a new and rapidly growing area of study in which researchers develop and test explanation methods (also called interpretability methods) that seek to shed some light on how black-box machine-learning models make predictions.

What are explanation methods?

At their most basic level, explanation methods are either global or local. A local explanation method focuses on explaining how the model made one specific prediction, while global explanations seek to describe the overall behavior of an entire model. This is often done by developing a separate, simpler (and hopefully understandable) model that mimics the larger, black-box model.

But because deep learning models work in fundamentally complex and nonlinear ways, developing an effective global explanation model is particularly challenging. This has led researchers to turn much of their recent focus onto local explanation methods instead, explains Yilun Zhou, a graduate student in the Interactive Robotics Group of the Computer Science and Artificial Intelligence Laboratory (CSAIL) who studies models, algorithms, and evaluations in interpretable machine learning.

The most popular types of local explanation methods fall into three broad categories.

The first and most widely used type of explanation method is known as feature attribution. Feature attribution methods show which features were most important when the model made a specific decision.

Features are the input variables that are fed to a machine-learning model and used in its prediction. When the data are tabular, features are drawn from the columns in a dataset (they are transformed using a variety of techniques so the model can process the raw data). For image-processing tasks, on the other hand, every pixel in an image is a feature. If a model predicts that an X-ray image shows cancer, for instance, the feature attribution method would highlight the pixels in that specific X-ray that were most important for the model's prediction.

Essentially, feature attribution methods show what the model pays the most attention to when it makes a prediction.

"Using this feature attribution explanation, you can check to see whether a spurious correlation is a concern. For instance, it will show if the pixels in a watermark are highlighted or if the pixels in an actual tumor are highlighted," says Zhou.

A second type of explanation method is known as a counterfactual explanation. Given an input and a model's prediction, these methods show how to change that input so it falls into another class. For instance, if a machine-learning model predicts that a borrower would be denied a loan, the counterfactual explanation shows what factors need to change so her loan application is accepted. Perhaps her credit score or income, both features used in the model's prediction, need to be higher for her to be approved.

"The good thing about this explanation method is it tells you exactly how you need to change the input to flip the decision, which could have practical usage. For someone who is applying for a mortgage and didn't get it, this explanation would tell them what they need to do to achieve their desired outcome," he says.

The third category of explanation methods are known as sample importance explanations. Unlike the others, this method requires access to the data that were used to train the model.

A sample importance explanation will show which training sample a model relied on most when it made a specific prediction; ideally, this is the most similar sample to the input data. This type of explanation is particularly useful if one observes a seemingly irrational prediction. There may have been a data entry error that affected a particular sample that was used to train the model. With this knowledge, one could fix that sample and retrain the model to improve its accuracy.

How are explanation methods used?

One motivation for developing these explanations is to perform quality assurance and debug the model. With more understanding of how features impact a model's decision, for instance, one could identify that a model is working incorrectly and intervene to fix the problem, or toss the model out and start over.

Another, more recent, area of research is exploring the use of machine-learning models to discover scientific patterns that humans haven't uncovered before. For instance, a cancer diagnosing model that outperforms clinicians could be faulty, or it could actually be picking up on some hidden patterns in an X-ray image that represent an early pathological pathway for cancer that were either unknown to human doctors or thought to be irrelevant, Zhou says.

It's still very early days for that area of research, however.

Words of warning

While explanation methods can sometimes be useful for machine-learning practitioners when they are trying to catch bugs in their models or understand the inner-workings of a system, end-users should proceed with caution when trying to use them in practice, says Marzyeh Ghassemi, an assistant professor and head of the Healthy ML Group in CSAIL.

As machine learning has been adopted in more disciplines, from health care to education, explanation methods are being used to help decision makers better understand a model's predictions so they know when to trust the model and use its guidance in practice. But Ghassemi warns against using these methods in that way.

"We have found that explanations make people, both experts and nonexperts, overconfident in the ability or the advice of a specific recommendation system. I think it is very important for humans not to

turn off that internal circuitry asking, 'let me question the advice that I am
given,'" she says.

Scientists know explanations make people over-confident based on other recent work, she adds, citing some recent studies by Microsoft researchers.

Far from a silver bullet, explanation methods have their share of problems. For one, Ghassemi's recent research has shown that explanation methods can perpetuate biases and lead to worse outcomes for people from disadvantaged groups.

Another pitfall of explanation methods is that it is often impossible to tell if the explanation method is correct in the first place. One would need to compare the explanations to the actual model, but since the user doesn't know how the model works, this is circular logic, Zhou says.

He and other researchers are working on improving explanation methods so they are more faithful to the actual model's predictions, but Zhou cautions that, even the best explanation should be taken with a grain of salt.

"In addition, people generally perceive these models to be human-like decision makers, and we are prone to overgeneralization. We need to calm people down and hold them back to really make sure that the generalized model understanding they build from these local explanations are balanced," he adds.

Zhou's most recent research seeks to do just that.

What's next for machine-learning explanation

methods?

Rather than focusing on providing explanations, Ghassemi argues that more effort needs to be done by the research community to study how information is presented to decision makers so they understand it, and more regulation needs to be put in place to ensure machine-learning models are used responsibly in practice. Better explanation methods alone aren't the answer.

"I have been excited to see that there is a lot more recognition, even in industry, that we can't just take this information and make a pretty dashboard and assume people will perform better with that. You need to have measurable improvements in action, and I'm hoping that leads to real guidelines about improving the way we display information in these deeply technical fields, like medicine," she says.

And in addition to new work focused on improving explanations, Zhou expects to see more research related to explanation methods for specific use cases, such as model debugging, scientific discovery, fairness auditing, and safety assurance. By identifying fine-grained characteristics of explanation methods and the requirements of different use cases, researchers could establish a theory that would match explanations with specific scenarios, which could help overcome some of the pitfalls that come from using them in real-world scenarios.

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

Citation: How to tell if artificial intelligence is working the way we want it to (2022, July 22) retrieved 27 April 2024 from https://techxplore.com/news/2022-07-artificial-intelligence.html