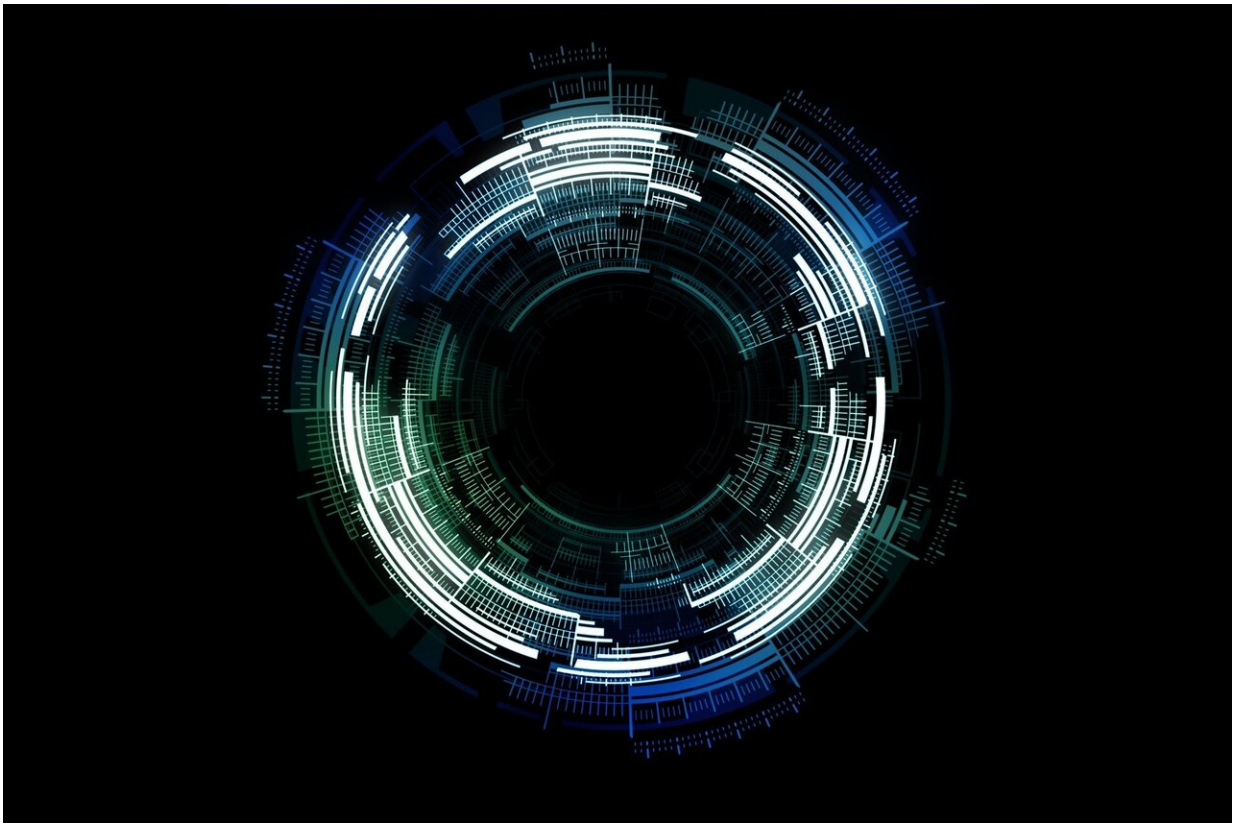
How well do explanation methods for machine-learning models work?

January 19 2022, by Adam Zewe



Credit: Pixabay/CC0 Public Domain

Imagine a team of physicians using a neural network to detect cancer in mammogram images. Even if this machine-learning model seems to be performing well, it might be focusing on image features that are

accidentally correlated with tumors, like a watermark or timestamp, rather than actual signs of tumors.

To test these models, researchers use "feature-attribution methods," techniques that are supposed to tell them which parts of the image are the most important for the neural network's prediction. But what if the attribution method misses features that are important to the model? Since the researchers don't know which features are important to begin with, they have no way of knowing that their evaluation method isn't effective.

To help solve this problem, MIT researchers have devised a process to modify the original data so they will be certain which features are actually important to the model. Then they use this modified dataset to evaluate whether feature-attribution methods can correctly identify those important features.

They find that even the most popular methods often miss the important features in an image, and some methods barely manage to perform as well as a random baseline. This could have major implications, especially if neural networks are applied in high-stakes situations like medical diagnoses. If the network isn't working properly, and attempts to catch such anomalies aren't working properly either, human experts may have no idea they are misled by the faulty model, explains lead author Yilun Zhou, an electrical engineering and computer science graduate student in the Computer Science and Artificial Intelligence Laboratory (CSAIL).

"All these methods are very widely used, especially in some really high-stakes scenarios, like detecting cancer from X-rays or CT scans. But these feature-attribution methods could be wrong in the first place. They may highlight something that doesn't correspond to the true feature the model is using to make a prediction, which we found to often be the

case. If you want to use these feature-attribution methods to justify that a model is working correctly, you better ensure the feature-attribution method itself is working correctly in the first place," he says.

Zhou wrote the paper with fellow EECS graduate student Serena Booth, Microsoft Research researcher Marco Tulio Ribeiro, and senior author Julie Shah, who is an MIT professor of aeronautics and astronautics and the director of the Interactive Robotics Group in CSAIL.

Focusing on features

In image classification, each pixel in an image is a feature that the neural network can use to make predictions, so there are literally millions of possible features it can focus on. If researchers want to design an algorithm to help aspiring photographers improve, for example, they could train a model to distinguish photos taken by professional photographers from those taken by casual tourists. This model could be used to assess how much the amateur photos resemble the professional ones, and even provide specific feedback on improvement. Researchers would want this model to focus on identifying artistic elements in professional photos during training, such as color space, composition, and postprocessing. But it just so happens that a professionally shot photo likely contains a watermark of the photographer's name, while few tourist photos have it, so the model could just take the shortcut of finding the watermark.

"Obviously, we don't want to tell aspiring photographers that a watermark is all you need for a successful career, so we want to make sure that our model focuses on the artistic features instead of the watermark presence. It is tempting to use feature attribution methods to analyze our model, but at the end of the day, there is no guarantee that they work correctly, since the model could use artistic features, the watermark, or any other features," Zhou says.

"We don't know what those spurious correlations in the dataset are. There could be so many different things that might be completely imperceptible to a person, like the resolution of an image," Booth adds. "Even if it is not perceptible to us, a neural network can likely pull out those features and use them to classify. That is the underlying problem. We don't understand our datasets that well, but it is also impossible to understand our datasets that well."

The researchers modified the dataset to weaken all the correlations between the original image and the data labels, which guarantees that none of the original features will be important anymore.

Then, they add a new feature to the image that is so obvious the neural network has to focus on it to make its prediction, like bright rectangles of different colors for different image classes.

"We can confidently assert that any model achieving really high confidence has to focus on that colored rectangle that we put in. Then we can see if all these feature-attribution methods rush to highlight that location rather than everything else," Zhou says.

"Especially alarming" results

They applied this technique to a number of different feature-attribution methods. For image classifications, these methods produce what is known as a saliency map, which shows the concentration of important features spread across the entire image. For instance, if the neural network is classifying images of birds, the saliency map might show that 80 percent of the important features are concentrated around the bird's beak.

After removing all the correlations in the image data, they manipulated the photos in several ways, such as blurring parts of the image, adjusting

the brightness, or adding a watermark. If the feature-attribution method is working correctly, nearly 100 percent of the important features should be located around the area the researchers manipulated.

The results were not encouraging. None of the feature-attribution methods got close to the 100 percent goal, most barely reached a random baseline level of 50 percent, and some even performed worse than the baseline in some instances. So, even though the new feature is the only one the model could use to make a prediction, the feature-attribution methods sometimes fail to pick that up.

"None of these methods seem to be very reliable, across all different types of spurious correlations. This is especially alarming because, in natural datasets, we don't know which of those spurious correlations might apply," Zhou says. "It could be all sorts of factors. We thought that we could trust these methods to tell us, but in our experiment, it seems really hard to trust them."

All feature-attribution methods they studied were better at detecting an anomaly than the absence of an anomaly. In other words, these methods could find a watermark more easily than they could identify that an image does not contain a watermark. So, in this case, it would be more difficult for humans to trust a model that gives a negative prediction.

The team's work shows that it is critical to test feature-attribution methods before applying them to a real-world model, especially in high-stakes situations.

"Researchers and practitioners may employ explanation techniques like feature-attribution methods to engender a person's trust in a model, but that trust is not founded unless the explanation technique is first rigorously evaluated," Shah says. "An explanation technique may be used to help calibrate a person's trust in a model, but it is equally important to

calibrate a person's trust in the explanations of the model."

Moving forward, the researchers want to use their evaluation procedure to study more subtle or realistic features that could lead to spurious correlations. Another area of work they want to explore is helping humans understand saliency maps so they can make better decisions based on a neural network's predictions.

More information: Do Feature Attribution Methods Correctly Attribute Features? arXiv:2104.14403 [cs.LG] arxiv.org/abs/2104.14403

Provided by Massachusetts Institute of Technology