New method compares machine-learning model's reasoning to that of a human

April 6 2022, by Adam Zewe

MIT researchers developed a method that helps a user understand a machine-learning model's reasoning, and how that reasoning compares to that of a human. Credit: Christine Daniloff, MIT

In machine learning, understanding why a model makes certain decisions

is often just as important as whether those decisions are correct. For instance, a machine-learning model might correctly predict that a skin lesion is cancerous, but it could have done so using an unrelated blip on a clinical photo.

While tools exist to help experts make sense of a model's reasoning, often these methods only provide insights on one decision at a time, and each must be manually evaluated. Models are commonly trained using millions of data inputs, making it almost impossible for a human to evaluate enough decisions to identify patterns.

Now, researchers at MIT and IBM Research have created a method that enables a user to aggregate, sort, and rank these individual explanations to rapidly analyze a machine-learning model's behavior. Their technique, called Shared Interest, incorporates quantifiable metrics that compare how well a model's reasoning matches that of a human.

Shared Interest could help a user easily uncover concerning trends in a model's decision-making—for example, perhaps the model often becomes confused by distracting, irrelevant features, like background objects in photos. Aggregating these insights could help the user quickly and quantitatively determine whether a model is trustworthy and ready to be deployed in a real-world situation.

"In developing Shared Interest, our goal is to be able to scale up this analysis process so that you could understand on a more global level what your model's behavior is," says lead author Angie Boggust, a graduate student in the Visualization Group of the Computer Science and Artificial Intelligence Laboratory (CSAIL).

Boggust wrote the paper with her advisor, Arvind Satyanarayan, an assistant professor of computer science who leads the Visualization Group, as well as Benjamin Hoover and senior author Hendrik Strobelt,

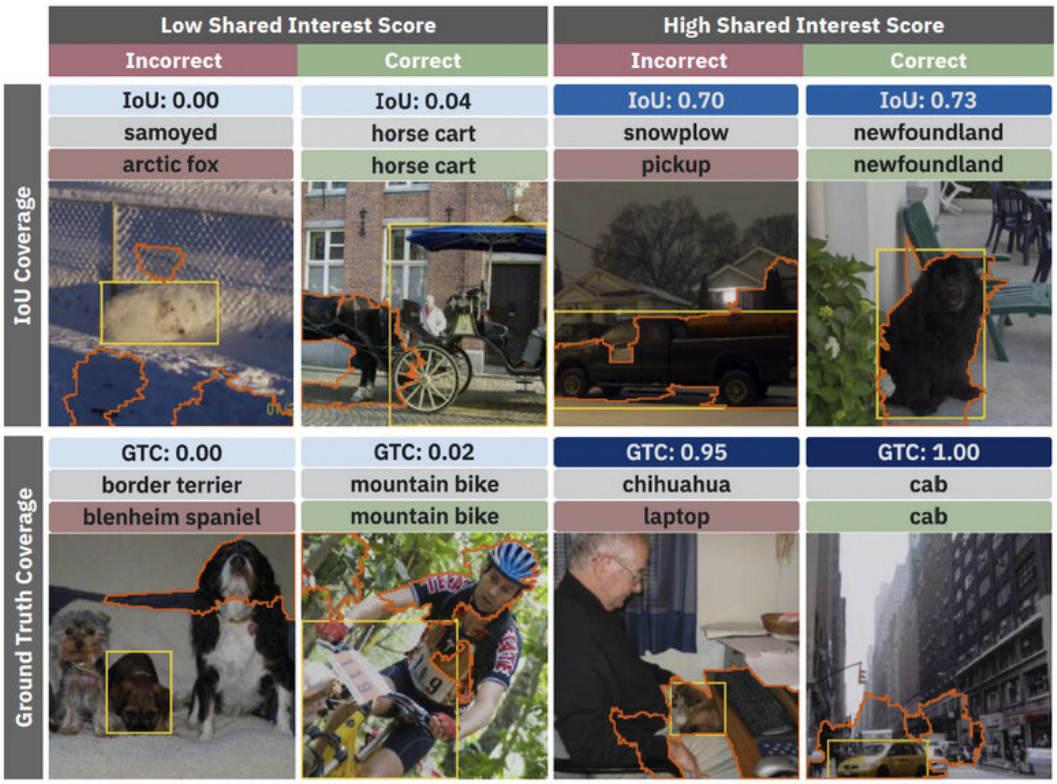
both of IBM Research. The paper will be presented at the Conference on Human Factors in Computing Systems.

Boggust began working on this project during a summer internship at IBM, under the mentorship of Strobelt. After returning to MIT, Boggust and Satyanarayan expanded on the project and continued the collaboration with Strobelt and Hoover, who helped deploy the case studies that show how the technique could be used in practice.

Human-AI alignment

Shared Interest leverages popular techniques that show how a machine-learning model made a specific decision, known as saliency methods. If the model is classifying images, saliency methods highlight areas of an image that are important to the model when it made its decision. These areas are visualized as a type of heatmap, called a saliency map, that is often overlaid on the original image. If the model classified the image as a dog, and the dog's head is highlighted, that means those pixels were important to the model when it decided the image contains a dog.

Shared Interest works by comparing saliency methods to ground-truth data. In an image dataset, ground-truth data are typically human-generated annotations that surround the relevant parts of each image. In the previous example, the box would surround the entire dog in the photo. When evaluating an image classification model, Shared Interest compares the model-generated saliency data and the human-generated ground-truth data for the same image to see how well they align.



Researchers developed a method that uses quantifiable metrics to compare how well a machine learning model's reasoning matches that of a human. This image shows the pixels in each picture that the model used to classify the image (surrounded by the orange line) and how that compares to the most important pixels, as defined by a human (surrounded by the yellow box). Credit: Massachusetts Institute of Technology

The technique uses several metrics to quantify that alignment (or misalignment) and then sorts a particular decision into one of eight categories. The categories run the gamut from perfectly human-aligned (the model makes a correct prediction and the highlighted area in the saliency map is identical to the human-generated box) to completely distracted (the model makes an incorrect prediction and does not use any image features found in the human-generated box).

"On one end of the spectrum, your model made the decision for the exact same reason a human did, and on the other end of the spectrum, your model and the human are making this decision for totally different reasons. By quantifying that for all the images in your dataset, you can use that quantification to sort through them," Boggust explains.

The technique works similarly with text-based data, where key words are highlighted instead of image regions.

Rapid analysis

The researchers used three case studies to show how Shared Interest could be useful to both nonexperts and machine-learning researchers.

In the first case study, they used Shared Interest to help a dermatologist determine if he should trust a machine-learning model designed to help diagnose cancer from photos of skin lesions. Shared Interest enabled the dermatologist to quickly see examples of the model's correct and incorrect predictions. Ultimately, the dermatologist decided he could not trust the model because it made too many predictions based on image artifacts, rather than actual lesions.

"The value here is that using Shared Interest, we are able to see these patterns emerge in our model's behavior. In about half an hour, the dermatologist was able to make a confident decision of whether or not to trust the model and whether or not to deploy it," Boggust says.

In the second case study, they worked with a machine-learning researcher to show how Shared Interest can evaluate a particular saliency method by revealing previously unknown pitfalls in the model. Their technique enabled the researcher to analyze thousands of correct and incorrect decisions in a fraction of the time required by typical manual methods.

In the third case study, they used Shared Interest to dive deeper into a specific image classification example. By manipulating the ground-truth area of the image, they were able to conduct a what-if analysis to see which image features were most important for particular predictions.

The researchers were impressed by how well Shared Interest performed in these case studies, but Boggust cautions that the technique is only as good as the saliency methods it is based upon. If those techniques contain bias or are inaccurate, then Shared Interest will inherit those limitations.

In the future, the researchers want to apply Shared Interest to different types of data, particularly tabular data which is used in medical records. They also want to use Shared Interest to help improve current saliency techniques. Boggust hopes this research inspires more work that seeks to quantify machine-learning model behavior in ways that make sense to humans.

More information: Angie Boggust, Benjamin Hoover, Arvind Satyanarayan, Hendrik Strobelt, Shared Interest: Measuring Human-AI Alignment to Identify Recurring Patterns in Model Behavior. arXiv:2107.09234v2 [cs.LG], arxiv.org/abs/2107.09234

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