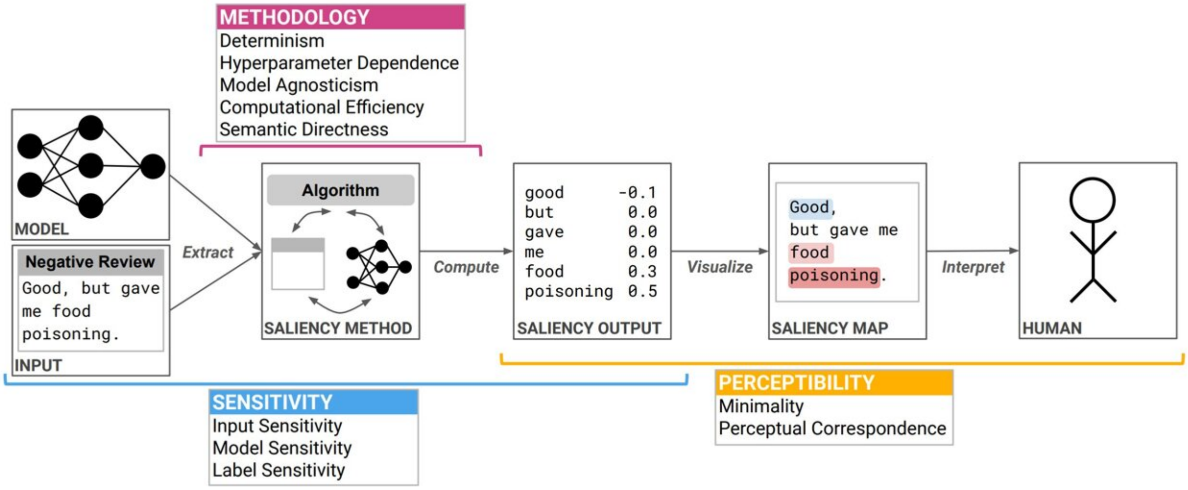
New tool helps people choose the right method for evaluating AI models

May 31 2023, by Adam Zewe



Saliency cards characterize saliency methods via ten user-centric attributes grouped into three categories corresponding to different phases of the interpretation process. Methodology attributes describe how the saliency is computed, sensitivity attributes express relationships between the saliency and its inputs, and perceptibility attributes measure human perception of the saliency. Credit: *Saliency Cards: A Framework to Characterize and Compare* (2023). Saliency Methods

When machine-learning models are deployed in real-world situations, perhaps to flag potential disease in X-rays for a radiologist to review, human users need to know when to trust the model's predictions.

But machine-learning models are so large and complex that even the scientists who design them don't understand exactly how the models make predictions. So, they create techniques known as saliency methods that seek to explain model behavior.

With new methods being released all the time, researchers from MIT and IBM Research created a tool to help users choose the best saliency method for their particular task. They developed saliency cards, which provide standardized documentation of how a method operates, including its strengths and weaknesses and explanations to help users interpret it correctly.

They hope that, armed with this information, users can deliberately select an appropriate saliency method for both the type of machine-learning model they are using and the task that model is performing, explains co-lead author Angie Boggust, a graduate student in electrical engineering and computer science at MIT and member of the Visualization Group of the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL).

Interviews with AI researchers and experts from other fields revealed that the cards help people quickly conduct a side-by-side comparison of different methods and pick a task-appropriate technique. Choosing the right method gives users a more accurate picture of how their model is behaving, so they are better equipped to correctly interpret its predictions.

"Saliency cards are designed to give a quick, glanceable summary of a saliency method and also break it down into the most critical, human-centric attributes. They are really designed for everyone, from machine-learning researchers to lay users who are trying to understand which method to use and choose one for the first time," says Boggust.

Joining Boggust on the paper are co-lead author Harini Suresh, an MIT postdoc; Hendrik Strobelt, a senior research scientist at IBM Research; John Guttag, the Dugald C. Jackson Professor of Computer Science and Electrical Engineering at MIT; and senior author Arvind Satyanarayan, associate professor of computer science at MIT who leads the Visualization Group in CSAIL. The research will be presented at the ACM Conference on Fairness, Accountability, and Transparency.

Picking the right method

The researchers have previously evaluated saliency methods using the notion of faithfulness. In this context, faithfulness captures how accurately a method reflects a model's decision-making process.

But faithfulness is not black-and-white, Boggust explains. A method might perform well under one test of faithfulness, but fail another. With so many saliency methods, and so many possible evaluations, users often settle on a method because it is popular or a colleague has used it.

However, picking the "wrong" method can have serious consequences. For instance, one saliency method, known as integrated gradients, compares the importance of features in an image to a meaningless baseline. The features with the largest importance over the baseline are most meaningful to the model's prediction. This method typically uses all 0s as the baseline, but if applied to images, all 0s equates to the color black.

"It will tell you that any black pixels in your image aren't important, even if they are, because they are identical to that meaningless baseline. This could be a big deal if you are looking at X-rays since black could be meaningful to clinicians," says Boggust.

Saliency cards can help users avoid these types of problems by summarizing how a saliency method works in terms of 10 user-focused attributes. The attributes capture the way saliency is calculated, the relationship between the saliency method and the model, and how a user perceives its outputs.

For example, one attribute is hyperparameter dependence, which measures how sensitive that saliency method is to user-specified parameters. A saliency card for integrated gradients would describe its parameters and how they affect its performance. With the card, a user could quickly see that the default parameters—a baseline of all 0s—might generate misleading results when evaluating X-rays.

The cards could also be useful for scientists by exposing gaps in the research space. For instance, the MIT researchers were unable to identify a saliency method that was computationally efficient, but could also be applied to any machine-learning model.

"Can we fill that gap? Is there a saliency method that can do both things? Or maybe these two ideas are theoretically in conflict with one another," Boggust says.

Showing their cards

Once they had created several cards, the team conducted a user study with eight domain experts, from computer scientists to a radiologist who was unfamiliar with machine learning. During interviews, all participants said the concise descriptions helped them prioritize attributes and compare methods. And even though he was unfamiliar with machine learning, the radiologist was able to understand the cards and use them to take part in the process of choosing a saliency method, Boggust says.

The interviews also revealed a few surprises. Researchers often expect that clinicians want a method that is sharp, meaning it focuses on a particular object in a medical image. But the clinician in this study actually preferred some noise in medical images to help them attenuate uncertainty.

"As we broke it down into these different attributes and asked people, not a single person had the same priorities as anyone else in the study, even when they were in the same role," she says.

Moving forward, the researchers want to explore some of the more under-evaluated attributes and perhaps design task-specific saliency methods. They also want to develop a better understanding of how people perceive saliency method outputs, which could lead to better visualizations. In addition, they are hosting their work on a public repository so others can provide feedback that will drive future work, Boggust says.

"We are really hopeful that these will be living documents that grow as new saliency methods and evaluations are developed. In the end, this is really just the start of a larger conversation around what the attributes of a saliency method are and how those play into different tasks," she says.

More information: Conference: facctconference.org/

Paper: vis.csail.mit.edu/pubs/saliency-cards.pdf

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: New tool helps people choose the right method for evaluating AI models (2023, May

31) retrieved 18 April 2024 from https://techxplore.com/news/2023-05-tool-people-method-ai.html