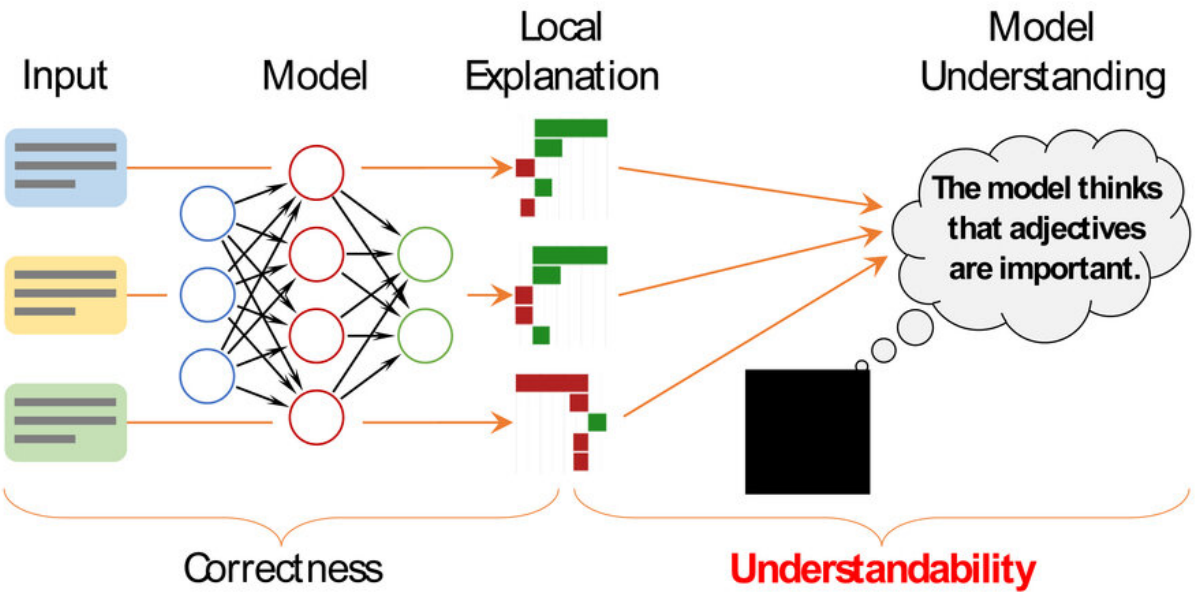
Framework to describe individual machine-learning model decisions

May 5 2022, by Adam Zewe



Researchers use local explanation methods to try and understand how machine learning models make decisions. Even if these explanations are correct, they don't do any good if humans can't understand what they mean. MIT researchers have now developed a mathematical framework to quantify and evaluate the understandability of an explanation. Credit: Massachusetts Institute of Technology

Modern machine-learning models, such as neural networks, are often referred to as "black boxes" because they are so complex that even the researchers who design them can't fully understand how they make predictions.

To provide some insights, researchers use explanation methods that seek to describe individual model decisions. For example, they may highlight words in a movie review that influenced the model's decision that the review was positive.

But these explanation methods don't do any good if humans can't easily understand them, or even misunderstand them. So, MIT researchers created a mathematical framework to formally quantify and evaluate the understandability of explanations for machine-learning models. This can help pinpoint insights about model behavior that might be missed if the researcher is only evaluating a handful of individual explanations to try to understand the entire model.

"With this framework, we can have a very clear picture of not only what we know about the model from these local explanations, but more importantly what we don't know about it," says Yilun Zhou, an electrical engineering and computer science graduate student in the Computer Science and Artificial Intelligence Laboratory (CSAIL) and lead author of a paper presenting this framework.

Zhou's co-authors include Marco Tulio Ribeiro, a senior researcher at Microsoft Research, and senior author Julie Shah, a professor of aeronautics and astronautics and the director of the Interactive Robotics Group in CSAIL. The research will be presented at the Conference of the North American Chapter of the Association for Computational Linguistics.

Understanding local explanations

One way to understand a machine-learning model is to find another model that mimics its predictions but uses transparent reasoning patterns. However, recent neural network models are so complex that this technique usually fails. Instead, researchers resort to using local explanations that focus on individual inputs. Often, these explanations highlight words in the text to signify their importance to one prediction made by the model.

Implicitly, people then generalize these local explanations to overall model behavior. Someone may see that a local explanation method highlighted positive words (like "memorable," "flawless," or "charming") as being the most influential when the model decided a movie review had a positive sentiment. They are then likely to assume that all positive words make positive contributions to a model's predictions, but that might not always be the case, Zhou says.

The researchers developed a framework, known as ExSum (short for explanation summary), that formalizes those types of claims into rules that can be tested using quantifiable metrics. ExSum evaluates a rule on an entire dataset, rather than just the single instance for which it is constructed.

Using a graphical user interface, an individual writes rules that can then be tweaked, tuned, and evaluated. For example, when studying a model that learns to classify movie reviews as positive or negative, one might write a rule that says "negation words have negative saliency," which means that words like "not," "no," and "nothing" contribute negatively to the sentiment of movie reviews.

Using ExSum, the user can see if that rule holds up using three specific metrics: coverage, validity, and sharpness. Coverage measures how broadly applicable the rule is across the entire dataset. Validity highlights the percentage of individual examples that agree with the rule. Sharpness

describes how precise the rule is; a highly valid rule could be so generic that it isn't useful for understanding the model.

Testing assumptions

If a researcher seeks a deeper understanding of how her model is behaving, she can use ExSum to test specific assumptions, Zhou says.

If she suspects her model is discriminative in terms of gender, she could create rules to say that male pronouns have a positive contribution and female pronouns have a negative contribution. If these rules have high validity, it means they are true overall and the model is likely biased.

ExSum can also reveal unexpected information about a model's behavior. For example, when evaluating the movie review classifier, the researchers were surprised to find that negative words tend to have more pointed and sharper contributions to the model's decisions than positive words. This could be due to review writers trying to be polite and less blunt when criticizing a film, Zhou explains.

"To really confirm your understanding, you need to evaluate these claims much more rigorously on a lot of instances. This kind of understanding at this fine-grained level, to the best of our knowledge, has never been uncovered in previous works," he says.

"Going from local explanations to global understanding was a big gap in the literature. ExSum is a good first step at filling that gap," adds Ribeiro.

Extending the framework

In the future, Zhou hopes to build upon this work by extending the

notion of understandability to other criteria and explanation forms, like counterfactual explanations (which indicate how to modify an input to change the model prediction). For now, they focused on feature attribution methods, which describe the individual features a model used to make a decision (like the words in a movie review).

In addition, he wants to further enhance the framework and user interface so people can create rules faster. Writing rules can require hours of human involvement—and some level of human involvement is crucial because humans must ultimately be able to grasp the explanations—but AI assistance could streamline the process.

As he ponders the future of ExSum, Zhou hopes their work highlights a need to shift the way researchers think about machine-learning model explanations.

"Before this work, if you have a correct local explanation, you are done. You have achieved the holy grail of explaining your model. We are proposing this additional dimension of making sure these explanations are understandable. Understandability needs to be another metric for evaluating our explanations," says Zhou.

More information: Yilun Zhou, Marco Tulio Ribeiro, Julie Shah, ExSum: From Local Explanations to Model Understanding. arXiv:2205.00130v1 [cs.CL], arxiv.org/abs/2205.00130

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