New method peeks inside the 'black box' of artificial intelligence

November 1 2018

	Question	Answer
Original	What color is the flower?	yellow
Reduced input	**flower?**	**yellow**

A new method to decode the decision-making processes used by 'black box' machine learning algorithms works by finding the minimum input that will still

yield a correct answer. In this example, the researchers first presented an algorithm with a photo of a sunflower and asked 'What color is the flower?' This resulted in the correct answer, 'yellow.' The researchers found that they could get the same correct answer, with a similarly high degree of confidence, by asking the algorithm a single-word question: 'Flower?' Credit: Shi Feng/University of Maryland

Artificial intelligence—specifically, machine learning—is a part of daily life for computer and smartphone users. From autocorrecting typos to recommending new music, machine learning algorithms can help make life easier. They can also make mistakes.

It can be challenging for computer scientists to figure out what went wrong in such cases. This is because many machine learning algorithms learn from information and make their predictions inside a virtual "black box," leaving few clues for researchers to follow.

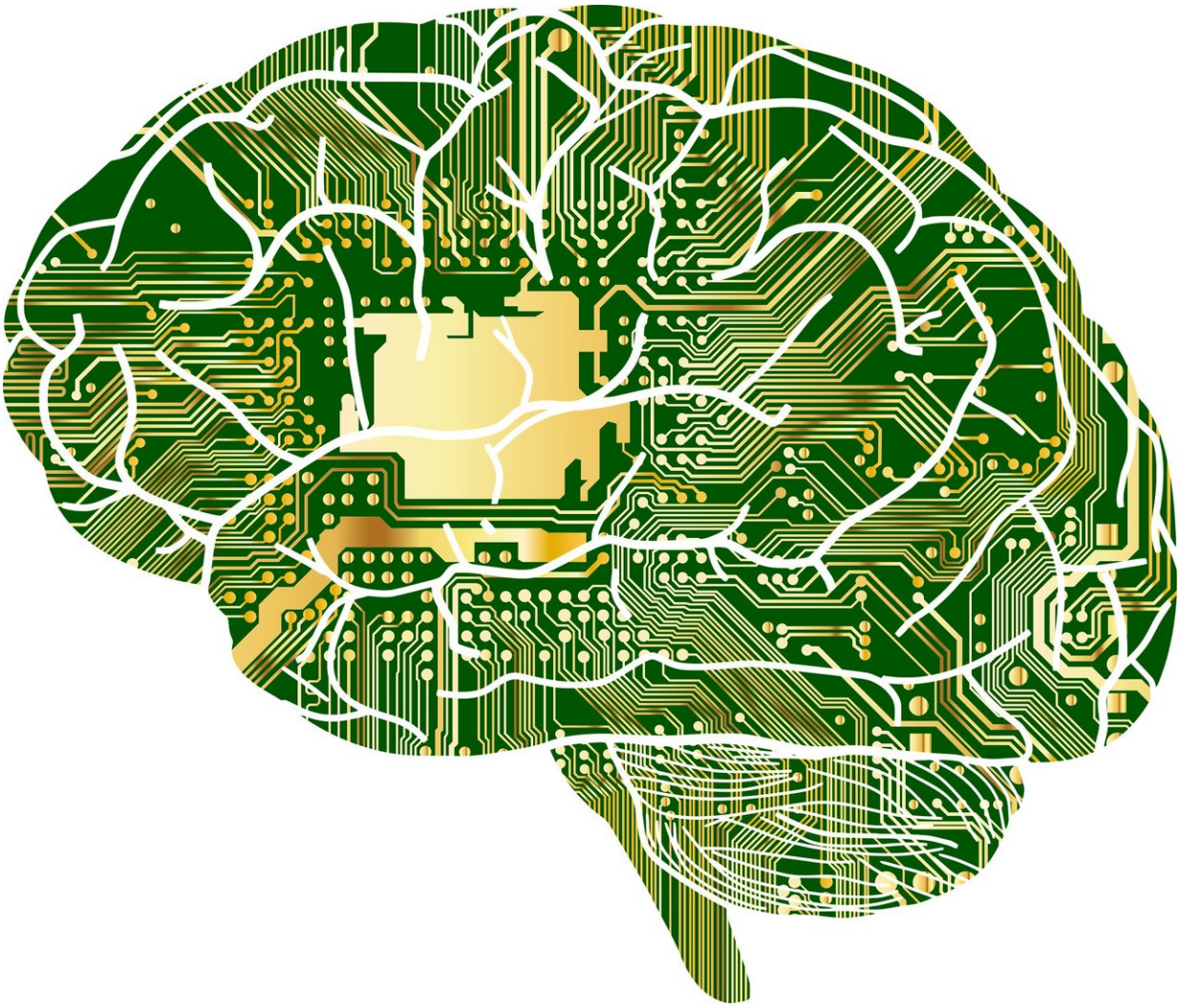
A group of computer scientists at the University of Maryland has developed a promising new approach for interpreting machine learning algorithms. Unlike previous efforts, which typically sought to "break" the algorithms by removing key words from inputs to yield the wrong answer, the UMD group instead reduced the inputs to the bare minimum required to yield the correct answer. On average, the researchers got the correct answer with an input of less than three words.

In some cases, the researchers' model algorithms provided the correct answer based on a single word. Frequently, the input word or phrase appeared to have little obvious connection to the answer, revealing important insights into how some algorithms react to specific language. Because many algorithms are programmed to give an answer no matter what—even when prompted by a nonsensical input—the results could

help computer scientists build more effective algorithms that can recognize their own limitations.

The researchers will present their work on November 4, 2018 at the 2018 Conference on Empirical Methods in Natural Language Processing.

"Black-box models do seem to work better than simpler models, such as decision trees, but even the people who wrote the initial code can't tell exactly what is happening," said Jordan Boyd-Graber, the senior author of the study and an associate professor of computer science at UMD. "When these models return incorrect or nonsensical answers, it's tough to figure out why. So instead, we tried to find the minimal input that would yield the correct result. The average input was about three words, but we could get it down to a single word in some cases."



Credit: CC0 Public Domain

In one example, the researchers entered a photo of a sunflower and the text-based question, "What color is the flower?" as inputs into a model algorithm. These inputs yielded the correct answer of "yellow." After rephrasing the question into several different shorter combinations of words, the researchers found that they could get the same answer with "flower?" as the only text input for the algorithm.

In another, more complex example, the researchers used the prompt, "In 1899, John Jacob Astor IV invested $100,000 for Tesla to further develop and produce a new lighting system. Instead, Tesla used the money to fund his Colorado Springs experiments."

They then asked the algorithm, "What did Tesla spend Astor's money on?" and received the correct answer, "Colorado Springs experiments." Reducing this input to the single word "did" yielded the same correct answer.

The work reveals important insights about the rules that machine learning algorithms apply to problem solving. Many real-world issues with algorithms result when an input that makes sense to humans results in a nonsensical answer. By showing that the opposite is also possible—that nonsensical inputs can also yield correct, sensible answers—Boyd-Graber and his colleagues demonstrate the need for algorithms that can recognize when they answer a nonsensical question with a high degree of confidence.

"The bottom line is that all this fancy machine learning stuff can actually be pretty stupid," said Boyd-Graber, who also has co-appointments at the University of Maryland Institute for Advanced Computer Studies (UMIACS) as well as UMD's College of Information Studies and Language Science Center. "When computer scientists train these models, we typically only show them real questions or real sentences. We don't show them nonsensical phrases or single words. The models don't know that they should be confused by these examples."

Most algorithms will force themselves to provide an answer, even with insufficient or conflicting data, according to Boyd-Graber. This could be at the heart of some of the incorrect or nonsensical outputs generated by machine learning algorithms—in model algorithms used for research, as well as real-world algorithms that help us by flagging spam email or

offering alternate driving directions. Understanding more about these errors could help computer scientists find solutions and build more reliable algorithms.

"We show that models can be trained to know that they should be confused," Boyd-Graber said. "Then they can just come right out and say, 'You've shown me something I can't understand.'"

More information: The research presentation, "Pathologies of Neural Models Make Interpretation Difficult," Shi Feng, Eric Wallace, Alvin Grissom II, Pedro Rodriguez, Mohit Iyyer, and Jordan Boyd-Graber, will be presented at the 2018 Conference on Empirical Methods in Natural Language Processing on November 4, 2018.

Provided by University of Maryland

Citation: New method peeks inside the 'black box' of artificial intelligence (2018, November 1) retrieved 27 April 2024 from https://techxplore.com/news/2018-11-method-peeks-black-artificial-intelligence.html

This document is subject to copyright. Apart from any fair dealing for the purpose of private study or research, no part may be reproduced without the written permission. The content is provided for information purposes only.