AI's mysterious 'black box' may not be so black

April 8 2024, by Victoria Skeidsvoll

"It is entirely possible to get more accurate information, and not just a 'hunch' about what happened or went wrong in an AI system. CIU has the potential to provide great opportunities for companies and their customers, but also for authorities and citizens," says Kary Främling, WASP Professor in Explainable AI at Umeå University. Credit: Mattias Pettersson

One of the pioneers of Explainable AI has developed an advanced model that explains how and why AI works. The model opens up AI's

mysterious "black box" and is available for virtually all AI systems.

"It can now be of great benefit to society and industry in understanding and explaining decisions made by AI, machine learning models and neural networks," says Kary Främling, Professor at the Department of Computer Science, Umeå University.

AI and machine learning are used by governments, health care, business, and industry. So-called deep learning methods can now diagnose patients in health care much faster than humans. But what is it that makes an AI system recommend one type of treatment and not another, and how does it come to its decisions?

"Explainable AI is an area that many people are interested in but few people know about or fully understand. The existing explanatory models are also not sufficiently comprehensible to the public," says Professor Främling, Head of The eXplainable Artificial Intelligence (XAI) team at The Department of Computing Science, Umeå University.

Främling has developed the CIU method (Contextual Importance and Utility approach) and finds it to be more efficient than other models. "Unfortunately, many scientists remained in a certain mindset, whereas I realized early on that these models were too limited. In the late 1990s, however, the time was not right, but I continued to develop the CIU method and today I can see that it was a valid choice in the long term," says Professor Främling.

Get a much more specific explanation

An AI system is a system where one or more inputs are given to an AI model or system, which then processes the information and produces one or more outputs. Främling uses his Ph.D. in France as an example.

The region where he lived wanted to identify the optimal location for the final storage of industrial waste. Thousands of sites were sorted out using machine learning and neural networks, and the choices were made considering several different categories. "But what were the criteria for deciding whether a site was good or not? Unfortunately, only a computer scientist like me could understand the reasoning of the AI system," says Främling.

The choice of location had to be justified, and consideration had to be given to both people and the environment. "You also have to explain it comprehensibly in different ways. Residents want one kind of information, while environmental authorities need another."

It was there and then that he became interested in creating a method of explanation. "For me, it's about ensuring that every one of us can understand the hospital's decision, the bank's response to a loan application or an authority's decision."

His CIU method allows you to study and explain the impact of changing one or more inputs—variables such as "age," "gender," "work" or "study"—on the final results. "CIU also allows you to calculate and explain each component and its impact on the results, as well as break down the input data into sub-sections. This means that you can get a much more specific explanation of, for example, why you didn't get the loan, or why you did," says Främling.

Provides understandable explanations

AI systems using neural networks, known as black box AI systems, were once considered impossible to explain. Therefore, so-called "surrogate models" were created in an attempt to imitate the operation of the actual AI system and analyze what it did. Explainable AI is still based on this idea. However, CIU does not create a surrogate model. Instead, it

analyzes the functioning of the AI model according to how outputs vary as a function of inputs.

"This provides information that can be translated into understandable explanations and concepts that we humans use to justify our decisions and actions, says Främling.

"It is entirely possible to get more accurate information, and not just a 'hunch' about what happened or went wrong in an AI system. CIU can provide great opportunities for companies and their customers, but also for authorities and citizens," says Främling.

CIU is implemented in the Python and R programming languages and its source code is publicly available on Github. CIU can also be installed as a library and, in principle, be integrated with any AI system. The model can even explain results from "classic" AI systems that do not use machine learning. The CIU can also be applied to time series and language models, but this is ongoing research.

Provided by Umea University

Citation: AI's mysterious 'black box' may not be so black (2024, April 8) retrieved 2 May 2024 from https://techxplore.com/news/2024-04-ai-mysterious-black.html