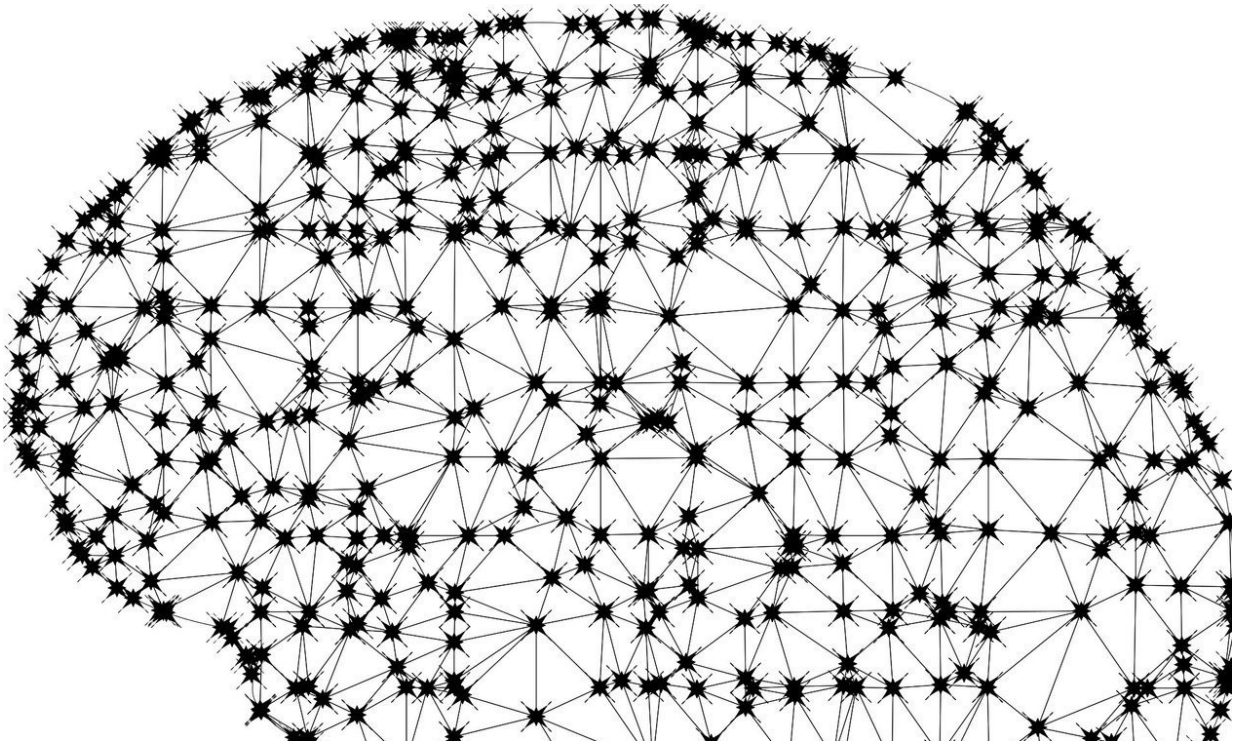
Taking machine thinking out of the black box

September 6 2018, by Anne Mcgovern



Credit: CC0 Public Domain

Software applications provide people with many kinds of automated decisions, such as identifying what an individual's credit risk is, informing a recruiter of which job candidate to hire, or determining whether someone is a threat to the public. In recent years, news headlines have warned of a future in which machines operate in the background of society, deciding the course of human lives while using

untrustworthy logic.

Part of this fear is derived from the obscure way in which many machine learning models operate. Known as black-box models, they are defined as systems in which the journey from input to output is next to impossible for even their developers to comprehend.

"As machine learning becomes ubiquitous and is used for applications with more serious consequences, there's a need for people to understand how it's making predictions so they'll trust it when it's doing more than serving up an advertisement," says Jonathan Su, a member of the technical staff in MIT Lincoln Laboratory's Informatics and Decision Support Group.

Currently, researchers either use post hoc techniques or an interpretable model such as a decision tree to explain how a black-box model reaches its conclusion. With post hoc techniques, researchers observe an algorithm's inputs and outputs and then try to construct an approximate explanation for what happened inside the black box. The issue with this method is that researchers can only guess at the inner workings, and the explanations can often be wrong. Decision trees, which map choices and their potential consequences in a tree-like construction, work nicely for categorical data whose features are meaningful, but these trees are not interpretable in important domains, such as computer vision and other complex data problems.

Su leads a team at the laboratory that is collaborating with Professor Cynthia Rudin at Duke University, along with Duke students Chaofan Chen, Oscar Li, and Alina Barnett, to research methods for replacing black-box models with prediction methods that are more transparent. Their project, called Adaptable Interpretable Machine Learning (AIM), focuses on two approaches: interpretable neural networks as well as adaptable and interpretable Bayesian rule lists (BRLs).

A neural network is a computing system composed of many interconnected processing elements. These networks are typically used for image analysis and object recognition. For instance, an algorithm can be taught to recognize whether a photograph includes a dog by first being shown photos of dogs. Researchers say the problem with these neural networks is that their functions are nonlinear and recursive, as well as complicated and confusing to humans, and the end result is that it is difficult to pinpoint what exactly the network has defined as "dogness" within the photos and what led it to that conclusion.

To address this problem, the team is developing what it calls "prototype neural networks." These are different from traditional neural networks in that they naturally encode explanations for each of their predictions by creating prototypes, which are particularly representative parts of an input image. These networks make their predictions based on the similarity of parts of the input image to each prototype.

As an example, if a network is tasked with identifying whether an image is a dog, cat, or horse, it would compare parts of the image to prototypes of important parts of each animal and use this information to make a prediction. A paper on this work: "This looks like that: deep learning for interpretable image recognition," was recently featured in an episode of the "Data Science at Home" podcast. A previous paper, "Deep Learning for Case-Based Reasoning through Prototypes: A Neural Network that Explains Its Predictions," used entire images as prototypes, rather than parts.

The other area the research team is investigating is BRLs, which are less-complicated, one-sided decision trees that are suitable for tabular data and often as accurate as other models. BRLs are made of a sequence of conditional statements that naturally form an interpretable model. For example, if blood pressure is high, then risk of heart disease is high. Su and colleagues are using properties of BRLs to enable users to indicate

which features are important for a prediction. They are also developing interactive BRLs, which can be adapted immediately when new data arrive rather than recalibrated from scratch on an ever-growing dataset.

Stephanie Carnell, a graduate student from the University of Florida and a summer intern in the Informatics and Decision Support Group, is applying the interactive BRLs from the AIM program to a project to help medical students become better at interviewing and diagnosing patients. Currently, medical students practice these skills by interviewing virtual patients and receiving a score on how much important diagnostic information they were able to uncover. But the score does not include an explanation of what, precisely, in the interview the students did to achieve their score. The AIM project hopes to change this.

"I can imagine that most medical students are pretty frustrated to receive a prediction regarding success without some concrete reason why," Carnell says. "The rule lists generated by AIM should be an ideal method for giving the students data-driven, understandable feedback."

The AIM program is part of ongoing research at the laboratory in human-systems engineering—or the practice of designing systems that are more compatible with how people think and function, such as understandable, rather than obscure, algorithms.

"The laboratory has the opportunity to be a global leader in bringing humans and technology together," says Hayley Reynolds, assistant leader of the Informatics and Decision Support Group. "We're on the cusp of huge advancements."

Melva James is another technical staff member in the Informatics and Decision Support Group involved in the AIM project. "We at the laboratory have developed Python implementations of both BRL and interactive BRLs," she says. "[We] are concurrently testing the output of

the BRL and interactive BRL implementations on different operating systems and hardware platforms to establish portability and reproducibility. We are also identifying additional practical applications of these algorithms."

Su explains: "We're hoping to build a new strategic capability for the laboratory—machine learning algorithms that people trust because they understand them."

More information: Deep Learning for Case-Based Reasoning through Prototypes: A Neural Network that Explains Its Predictions. arxiv.org/abs/1710.04806

This looks like that: deep learning for interpretable image recognition. arxiv.org/abs/1806.10574

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