Researchers develop 'explainable' artificial intelligence algorithm

April 1 2021, by Matthew Tierney

Heat-map images are used to evaluate the accuracy of a new explainable artificial intelligence algorithm that U of T and LG researchers developed to detect defects in LG's display screens. Credit: Mahesh Sudhakar

Researchers from the University of Toronto and LG AI Research have

developed an "explainable" artificial intelligence (XAI) algorithm that can help identify and eliminate defects in display screens.

The new algorithm, which outperformed comparable approaches on industry benchmarks, was developed through an ongoing AI research collaboration between LG and U of T that was expanded in 2019 with a focus on AI applications for businesses.

Researchers say the XAI algorithm could potentially be applied in other fields that require a window into how machine learning makes its decisions, including the interpretation of data from medical scans.

"Explainability and interpretability are about meeting the quality standards we set for ourselves as engineers and are demanded by the end user," says Kostas Plataniotis, a professor in the Edward S. Rogers Sr. department of electrical and computer engineering in the Faculty of Applied Science & Engineering. "With XAI, there's no 'one size fits all." You have to ask whom you're developing it for. Is it for another machine learning developer? Or is it for a doctor or lawyer?"

The research team also included recent U of T Engineering graduate Mahesh Sudhakar and master's candidate Sam Sattarzadeh, as well as researchers led by Jongseong Jang at LG AI Research Canada—part of the company's global research-and-development arm.

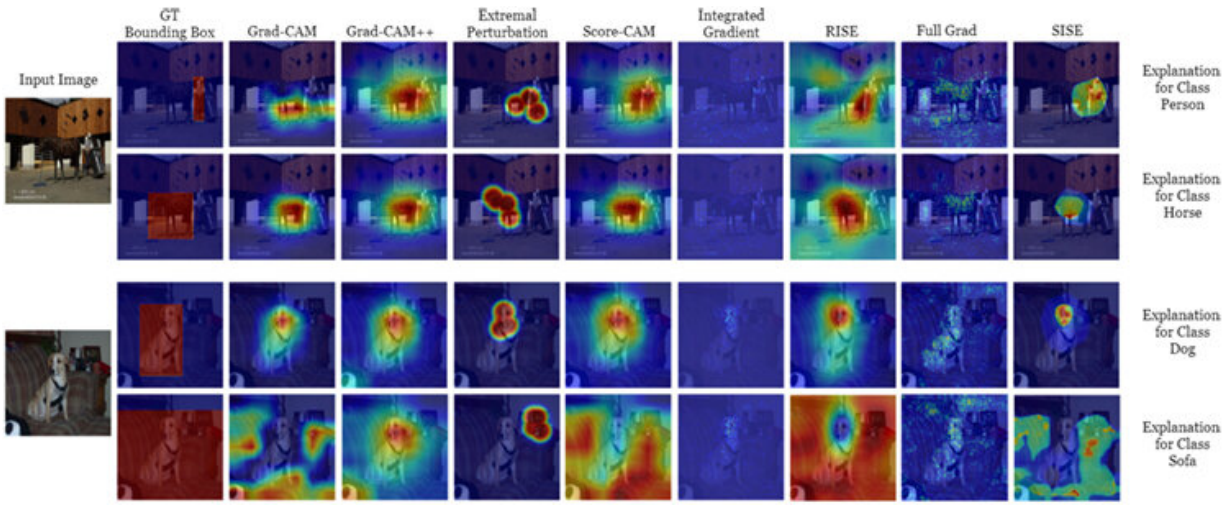
XAI is an emerging field that addresses issues with the 'black box' approach of machine learning strategies.

In a black box model, a computer might be given a set of training data in the form of millions of labeled images. By analyzing the data, the algorithm learns to associate certain features of the input (images) with certain outputs (labels). Eventually, it can correctly attach labels to images it has never seen before.

The machine decides for itself which aspects of the image to pay attention to and which to ignore, meaning its designers will never know exactly how it arrives at a result.

But such a "black box" model presents challenges when it's applied to areas such as health care, law and insurance.

"For example, a [machine learning] model might determine a patient has a 90 percent chance of having a tumor," says Sudhakar. "The consequences of acting on inaccurate or biased information are literally life or death. To fully understand and interpret the model's prediction, the doctor needs to know how the algorithm arrived at it."



Heat maps of industry benchmark images show a qualitative comparison of the team's XAI algorithm (SISE, far right) with other state-of-the-art XAI methods. Credit: Mahesh Sudhakar

In contrast to traditional machine learning, XAI is designed to be a "glass box" approach that makes the decision-making transparent. XAI

algorithms are run simultaneously with traditional algorithms to audit the validity and the level of their learning performance. The approach also provides opportunities to carry out debugging and find training efficiencies.

Sudhakar says that, broadly speaking, there are two methodologies to develop an XAI algorithm—each with advantages and drawbacks.

The first, known as back propagation, relies on the underlying AI architecture to quickly calculate how the network's prediction corresponds to its input. The second, known as perturbation, sacrifices some speed for accuracy and involves changing data inputs and tracking the corresponding outputs to determine the necessary compensation.

"Our partners at LG desired a new technology that combined the advantages of both," says Sudhakar. "They had an existing [machine learning] model that identified defective parts in LG products with displays, and our task was to improve the accuracy of the high-resolution heat maps of possible defects while maintaining an acceptable run time."

The team's resulting XAI algorithm, Semantic Input Sampling for Explanation (SISE), is described in a recent paper for the 35th AAAI Conference on Artificial Intelligence.

"We see potential in SISE for widespread application," says Plataniotis. "The problem and intent of the particular scenario will always require adjustments to the algorithm—but these heat maps or 'explanation maps' could be more easily interpreted by, for example, a medical professional."

"LG's goal in partnering with University of Toronto is to become a world leader in AI innovation," says Jang. "This first achievement in XAI speaks to our company's ongoing efforts to make contributions in

multiple areas, such as functionality of LG products, innovation of manufacturing, management of supply chain, efficiency of material discovery and others, using AI to enhance customer satisfaction."

Professor Deepa Kundur, chair of the electrical and computer engineering department, says successes like this are a good example of the value of collaborating with industry partners.

"When both sets of researchers come to the table with their respective points of view, it can often accelerate the problem-solving," Kundur says. "It is invaluable for graduate students to be exposed to this process."

While it was a challenge for the team to meet the aggressive accuracy and run-time targets within the year-long project—all while juggling Toronto/Seoul time zones and working under COVID-19 constraints—Sudhakar says the opportunity to generate a practical solution for a world-renowned manufacturer was well worth the effort.

"It was good for us to understand how, exactly, industry works," says Sudhakar. "LG's goals were ambitious, but we had very encouraging support from them, with feedback on ideas or analogies to explore. It was very exciting."

More information: Explaining Convolutional Neural Networks through Attribution-Based Input Sampling and Block-Wise Feature Aggregation. arXiv:2010.00672v2 [cs.CV] arxiv.org/abs/2010.00672

Provided by University of Toronto