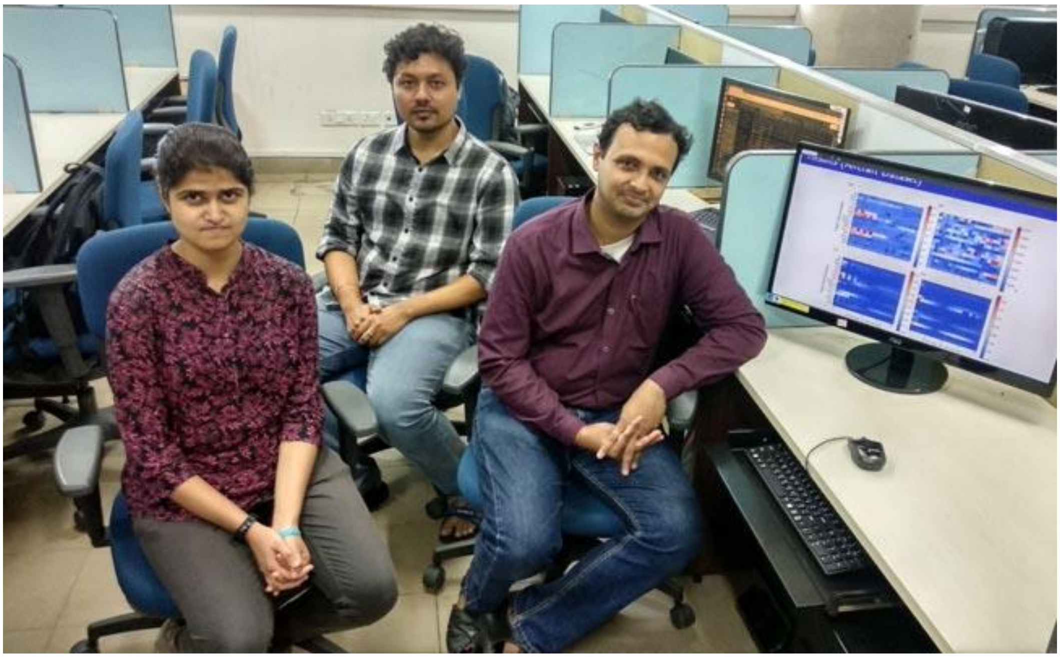
IIT Hyderabad Researchers Develop Method To Understand AI Models

IIT Hyderabad researchers have developed a method by which the inner workings of Artificial Intelligence models can be understood in terms of causal attributes.



IIT Hyderabad researchers have developed method to access insides of AI programs

NEW DELHI: IIT Hyderabad researchers have developed a method by which the inner workings of Artificial Intelligence models can be understood in terms of causal attributes.

'Artificial Neural Networks' (ANN) are AI models and programs that mimic the working of the human brain so that machines can learn to make decisions in a more human-like manner. Modern ANNs, often also called Deep Learning (DL), have increased tremendously in complexity such that machines can train themselves to process and learn from data that has been supplied to them as input, and almost match human performance in many tasks. However, how they arrive at decisions is unknown, making them less useful when the reason for decisions is necessary.

This work has been performed by Dr. Vineeth N. Balasubramanian, Associate Professor, Department of Computer Science and Engineering, IIT Hyderabad, and his students Mr. Aditya Chattopadhyay, Mr. Piyushi Manupriya, and Mr. Anirban Sarkar.

Their work was also recently published in the Proceedings of 36th International Conference on Machine Learning, which is one of the highest-rated conferences in the area of Artificial Intelligence and Machine Learning.

Speaking about this research, Dr. Vineeth Balasubramanian said, "The simplest applications that we know of Deep Learning (DL) is in machine translation, speech recognition or face detection. It enables voice-based control in consumer devices such as phones, tablets, television sets and hands-free speakers. New algorithms are being used in a variety of disciplines including engineering, finance, artificial perception and control and simulation. Much as the achievements have wowed everyone, there are challenges to be met."

A key bottleneck in accepting such Deep Learning models in real-life applications, especially risk-sensitive ones, is the 'interpretability problem.' The DL models, because of their complexity and multiple layers, become virtual black boxes that cannot be deciphered easily. Thus, when a problem arises in the running of the DL algorithm, troubleshooting becomes difficult, if not impossible, said Dr. Vineeth Balasubramanian.

The DL algorithms are trained on a limited amount of data that are most often different from real-world data. Furthermore, human error during training, and unnecessary correlations in data can result in errors that must be corrected, which becomes hard. "If treated as blackboxes, there is no way of knowing whether the model actually learned a concept or a high accuracy was just fortuitous," added Dr. Vineeth Balasubramanian.

The practical implications of the lack of transparency in DL models are that end-users can lose their trust over the system. There is thus a need for methods that can access the underbelly of the AI programs and unravel their structure and functions. The IIT Hyderabad team approached this problem with ANN architectures using causal inference with what is known in the field as a 'Structural Causal Model.'

Explaining this area of work, Dr. Balasubramanian said, "Thanks to our students' efforts and hard work, we have proposed a new method to compute the Average Causal Effect of an input neuron on an output neuron. It is important to understand which input parameter is 'causally' responsible for a given output; for example in the field of medicine, how does one know which patient attribute was causally responsible for the heart attack? Our (IIT Hyderabad researchers') method provides a tool to analyze such causal effects."

Transparency and understandability of the workings of DL models are gaining importance as discussions around the ethics of Artificial intelligence grows, added Dr. Balasubramanian on the importance of his team's work on 'explainable machine learning.' This makes sense given that the European Union General Data Protection Regulation (GDPR) regulation requires that an explanation must be provided if a machine learning model is used for any decisions made on its citizens, on any domain, be it banking, security or health.

The code developed by the IIT Hyderabad researchers to understand the workings of DL models, is available for free on Github. The research paper is also available in the public domain.

Source: NDTV
Date: 10/09/2019