UNIVERSITÉ DE GENÈVE

PRESS RELEASE

Geneva | 22 March 2023

NUS National University of Singapore

HUG Hôpitaux Universitaires Genève

Shining a light into the “black box” of AI

An international team led by UNIGE, HUG and NUS has developed an innovative method for evaluating AI interpretability methods, with the aim of deciphering the basis of AI reasoning and possible biases.

Researchers from the University of Geneva (UNIGE), the Geneva University Hospitals (HUG), and the National University of Singapore (NUS) have developed a novel method for evaluating the interpretability of artificial intelligence (AI) technologies, opening the door to greater transparency and trust in AI-driven diagnostic and predictive tools. The innovative approach sheds light on the opaque workings of so-called “black box” AI algorithms, helping users understand what influences the results produced by AI and whether the results can be trusted. This is especially important in situations that have significant impacts on the health and lives of people, such as using AI in medical applications. The research carries particular relevance in the context of the forthcoming European Union Artificial Intelligence Act which aims to regulate the development and use of AI within the EU. The findings have recently been published in the journal *Nature Machine Intelligence*.

Time series data - representing the evolution of information over time - is everywhere: for example in medicine, when recording heart activity with an electrocardiogram (ECG); in the study of earthquakes; tracking weather patterns; or in economics to monitor financial markets. This data can be modelled by AI technologies to build diagnostic or predictive tools.

The progress of AI and deep learning in particular - which consists of training a machine using these very large amounts of data with the aim of interpreting it and learning useful patterns - opens the pathway to increasingly accurate tools for diagnosis and prediction. Yet with no insight into how AI algorithms work or what influences their results, the “black box” nature of AI technology raises important questions over trustworthiness.

“The way these algorithms work is opaque, to say the least,” says Professor Christian Lovis, Director of the Department of Radiology and Medical Informatics at the UNIGE Faculty of Medicine and Head of the Division of Medical Information Science at the HUG, who co-directed this work. “Of course, the stakes, particularly financial, are extremely high. But how can we trust a machine without understanding the basis of its reasoning? These questions are essential, especially in sectors such as medicine, where AI-powered decisions can influence the health and even the lives of people; and finance, where they can lead to enormous loss of capital.”

High resolution pictures

Interpretability methods aim to answer these questions by deciphering why and how an AI reached a given decision, and the reasons behind it. “Knowing what elements tipped the scales in favour of or against a solution in a specific situation, thus allowing some transparency, increases the trust that can be placed in them,” says Assistant Professor Gianmarco Mengaldo, Director of the MathEXLab at the National University of Singapore’s College of Design and Engineering, who co-directed the work. “However, the current interpretability methods that are widely used in practical applications and industrial workflows provide tangibly different results when applied to the same task. This raises the important question: what interpretability method is correct, given that there should be a unique, correct answer? Hence, the evaluation of interpretability methods becomes as important as interpretability per se.”

Differentiating important from unimportant

Discriminating data is critical in developing interpretable AI technologies. For example, when an AI analyses images, it focuses on a few characteristic attributes. Doctoral student in Prof Lovis’ laboratory and first author of the study Hugues Turbé explains: “AI can, for example, differentiate between an image of a dog and an image of a cat. The same principle applies to analysing time sequences: the machine needs to be able to select elements - peaks that are more pronounced than others, for example - to base its reasoning on. With ECG signals, it means reconciling signals from the different electrodes to evaluate possible dissonances that would be a sign of a particular cardiac disease.”

Choosing an interpretability method among all available for a specific purpose is not easy. Different AI interpretability methods often produce very different results, even when applied on the same dataset and task. To address this challenge the researchers developed two new evaluation methods to help understand how the AI makes decisions: one for identifying the most relevant portions of a signal and another for evaluating their relative importance with regards to the final prediction. To evaluate interpretability, they hid a portion of the data to verify if it was relevant for the AI’s decision-making. However, this approach sometimes caused errors in the results. To correct for this, they trained the AI on an augmented dataset that includes hidden data which helped keep the data balanced and accurate. The team then created two ways to measure how well the interpretability methods worked, showing if the AI was using the right data to make decisions and if all the data was being considered fairly. “Overall our method aims to evaluate the model that will actually be used within its operational domain, thus ensuring its reliability,” explains Hugues Turbé.

To further their research, the team has developed a synthetic dataset, which they have made available to the scientific community, to easily evaluate any new AI aimed at interpreting temporal sequences.

The future of medical applications

Going forward, the team now plan to test their method in a clinical setting, where apprehension about AI remains widespread. “Building confidence in the evaluation of AIs is a key step towards their adoption in clinical settings,” explains Dr. Mina Bjelogrlic, who heads the Machine Learning team in Prof Lovis’ Division and is the second author of this study. “Our study focuses on the evaluation of AIs based on time series, but the same methodology could be applied to AIs based on other modalities used in medicine, such as images or text.”

contact

Christian Lovis
Full Professor
Director
Department of Radiology and Medical Informatics
Faculty of Medicine
UNIGE

Chief Medical Officer
Division of Medical Informatics science
HUG

+41 22 372 88 83

Christian.Lovis@unige.ch

Gianmarco Mengaldo
Assistant Professor
Director, MathEXLab
Dept of Mechanical Engineering

College of Design and Engineering
National University of Singapore

+65 6516 8023
mpegim@nus.edu.sg

Hugues Turbé
PhD student
Department of Radiology and Medical Informatics
Faculty of Medicine
UNIGE

+41 22 37 90815
Hugues.Turbe@unige.ch

DOI: 10.1038/s42256-023-00620-w

UNIVERSITÉ DE GENÈVE
Communication Department
24 rue du Général-Dufour
CH-1211 Geneva 4

Tel. +41 22 379 77 17

media@unige.ch
www.unige.ch