TechXplore

Common approach to demystify black box AI not ready for prime time

October 10 2022

Credit: Pixabay/CC0 Public Domain

Artificial intelligence models that interpret medical images hold the promise to enhance clinicians' ability to make accurate and timely diagnoses, while also lessening workload by allowing busy physicians to focus on critical cases and delegate rote tasks to AI.

But AI models that lack transparency about how and why a diagnosis is made can be problematic. This opaque reasoning—also known "black box" AI—can diminish clinician trust in the reliability of the AI tool and thus discourage its use. This lack of transparency could also mislead clinicians into over-trusting the tool's interpretation.

In the realm of medical imaging, one way to create more understandable AI models and to demystify AI decision-making have been saliency assessments—an approach that uses heat maps to pinpoint whether the tool is correctly focusing only on the relevant pieces of a given image or homing in on irrelevant parts of it.

Heat maps work by highlighting areas on an image that influenced the AI model's interpretation. This could help human physicians see whether the AI model focuses on the same areas as they do or is mistakenly focusing on irrelevant spots on an image.

But a new study, published in *Nature Machine Intelligence* on Oct. 10, shows that for all their promise, saliency heat maps may not be yet ready for prime time.

The analysis, led by Harvard Medical School investigator Pranav Rajpurkar, Matthew Lungren of Stanford, and Adriel Saporta of New York University, quantified the validity of seven widely used saliency methods to determine how reliably and accurately they could identify pathologies associated with 10 conditions commonly diagnosed on X-ray, such as lung lesions, pleural effusion, edema, or enlarged heart structures. To ascertain performance, the researchers compared the tools' performance against human expert judgment.

In the final analysis, tools using saliency-based heat maps consistently underperformed in image assessment and in their ability to spot pathological lesions, compared with human radiologists.

The work represents the first comparative analysis between saliency maps and human expert performance in the evaluation of multiple X-ray pathologies. The study also offers a granular understanding of whether and how certain pathological characteristics on an image might affect AI tool performance.

The saliency-map feature is already used as a quality assurance tool by clinical practices that employ AI to interpret computer-aided detection methods, such as reading chest X-rays. But in light of the new findings, this feature should be applied with caution and a healthy dose of skepticism, the researchers said.

"Our analysis shows that saliency maps are not yet reliable enough to validate individual clinical decisions made by an AI model," said Rajpurkar, who is an assistant professor of biomedical informatics at HMS. "We identified important limitations that raise serious safety concerns for use in current practice."

The researchers caution that because of the important limitations identified in the study, saliency-based heat maps should be further refined before they are widely adopted in clinical AI models.

The team's full codebase, data, and analysis are open and available to all interested in studying this important aspect of clinical machine learning in medical imaging applications.

More information: Pranav Rajpurkar, Benchmarking saliency methods for chest X-ray interpretation, *Nature Machine Intelligence* (2022). DOI: 10.1038/s42256-022-00536-x

Provided by Harvard Medical School

Citation: Common approach to demystify black box AI not ready for prime time (2022, October 10) retrieved 25 April 2024 from https://techxplore.com/news/2022-10-common-approach-demystify-black-ai.html