UNIVERSITÉ DE GENÈVE

NUS National University of Singapore

Hôpitaux Universitaires Genève

COMMUNIQUÉ DE PRESSE

Genève | 22 mars 2023

Dans la boîte noire des intelligences artificielles

Une équipe internationale conduite par l'UNIGE, les HUG et la NUS a développé une méthode innovante d'évaluation des IA afin de décrypter les bases de leur raisonnement et identifier leurs biais éventuels.

Illustrations haute définition

Des scientifiques de l'Université de Genève (UNIGE), des Hôpitaux universitaires de Genève (HUG) et de l'Université nationale de Singapour (NUS) ont mis au point une nouvelle méthode pour évaluer l'interprétabilité des technologies d'intelligence artificielle (IA), ouvrant la voie à une plus grande transparence des outils de diagnostic et de prédiction basés sur l'IA. Cette approche met en lumière le fonctionnement opaque des «boîtes noires» des algorithmes d'IA et permet de comprendre ce qui influence les résultats produits par l'IA, facilitant ainsi l'évaluation de leur fiabilité. C'est particulièrement important lorsque la santé et la vie des personnes sont en jeu, comme dans le cas des applications médicales. Cette recherche est d'autant plus pertinente dans le contexte de la future loi de l'Union européenne sur l'intelligence artificielle, qui vise à réguler le développement et l'usage des IA dans l'UE. Ces résultats sont publiés dans la revue *Nature Machine Intelligence*.

Les données de séries temporelles, qui représentent l'évolution de l'information dans le temps, sont omniprésentes: en médecine, lors de l'enregistrement de l'activité cardiaque à l'aide d'un électrocardiogramme (ECG), pour l'étude des tremblements de terre, pour suivre les schémas climatiques ou encore pour surveiller les marchés financiers. Ces données peuvent être modélisées par des technologies d'IA pour construire des outils diagnostiques ou prédictifs.

Les progrès du *deep learning*, qui consiste à entraîner une machine avec un très grand nombre de données afin qu'elle les interprète et en apprenne des motifs utiles, ouvrent des possibilités immenses pour des diagnostics et des prédictions de plus en plus précis. Cependant, comme on ignore comment fonctionnent les algorithmes d'IA, de même que ce qui influence leurs résultats, la «boîte noire» de la technologie de l'IA soulève d'importantes questions de confiance.

«Le fonctionnement de ces algorithmes est pour le moins opaque», souligne le professeur Christian Lovis, directeur du Département de radiologie et informatique médicale de la Faculté de médecine de l'UNIGE et médecin-chef du Service des sciences de l'information médicale des HUG, qui a co-dirigé ces travaux. «Certes, les enjeux, financiers notamment, sont énormes. Mais comment faire confiance à une machine sans comprendre les bases de son raisonnement? Ces questions sont essentielles lorsque les décisions basées sur des IA peuvent influer des décisions critiques, comme dans les applications médicales et leurs effets sur les patientes et les patients, ou dans le secteur financier, où elles peuvent mener à d'importantes pertes financières.»

L'objectif des méthodes d'interprétabilité est d'apporter une réponse à ces considérations en décryptant comment et pourquoi une IA est parvenue à une décision donnée et les raisons qui la sous-tendent. «Savoir quels sont les éléments qui ont penché en faveur ou en défaveur d'une solution dans une situation précise et ainsi apporter un peu de transparence à ces outils permet d'augmenter la confiance que l'on peut leur accorder», indique le professeur assistant Gianmarco Mengaldo, directeur du MathEXLab à la National University of Singapore, dernier auteur de cette étude qui a co-dirigé ce travail. «Cependant, les méthodes d'interprétabilité actuelles, largement utilisées dans les applications pratiques et les flux de travail industriels, fournissent des résultats sensiblement différents lorsqu'elles sont appliquées à la même tâche. Cela soulève une question importante: quelle méthode d'interprétabilité est correcte, étant donné qu'il devrait y avoir une seule réponse correcte? L'évaluation des méthodes d'interprétabilité devient donc aussi importante que l'interprétabilité elle-même.»

Différencier l'important de l'inutile

Identifier les données importantes est essentiel pour développer des technologies d'IA interprétables. Par exemple, lorsqu'une IA analyse des images, elle se concentre sur quelques attributs caractéristiques. «C'est ainsi qu'elle peut différencier une image de chien d'une image de chat. Le même principe s'applique pour analyser des séquences temporelles: il faut que la machine puisse sélectionner les éléments — des pics plus prononcés que d'autres, par exemple — sur lesquels baser son raisonnement. Avec des signaux d'ECG, il s'agira de réconcilier les signaux des différentes électrodes afin d'évaluer d'éventuelles dissonances qui seraient le signe de telle ou telle maladie cardiaque», explique Hugues Turbé, doctorant dans le laboratoire de Christian Lovis et premier auteur de l'étude.

Le choix d'une méthode d'interprétabilité parmi toutes celles disponibles pour un usage spécifique n'est pas chose aisée. En effet, différentes méthodes d'interprétabilité appliquées aux mêmes données et les mêmes tâches produisent souvent des résultats d'interprétabilité très différents.

Pour relever ce défi, les chercheurs et chercheuses ont mis au point deux nouvelles méthodes d'évaluation afin de mieux comprendre comment l'IA prend ses décisions: l'une pour identifier les parties les plus pertinentes d'un signal et l'autre pour évaluer leur importance relative par rapport à la prédiction finale. Pour évaluer l'interprétabilité, l'équipe a caché une partie des données afin de vérifier si elle était pertinente pour la prise de décision de l'IA. Cependant, cette approche peut parfois entraîner des erreurs dans les résultats. Pour y remédier, les scientifiques ont entraîné l'IA sur un ensemble de données enrichi comprenant des données cachées, ce qui a permis de maintenir l'équilibre et la précision des données. Ils et elles ont ensuite créé deux moyens de mesurer l'efficacité des méthodes d'interprétabilité, en montrant si l'IA utilisait les bonnes données pour prendre des

décisions et si toutes les données étaient prises en compte de manière équitable. «Notre méthode permet donc d'évaluer le modèle qui sera réellement utilisé dans son contexte opérationnel, et en assure ainsi la fiabilité», détaille Hugues Turbé.

L'équipe de recherche a également développé un jeu de données synthétique — mis à la disposition de la communauté scientifique — qui permet d'évaluer facilement toute nouvelle IA visant à interpréter des séquences temporelles.

Le futur des applications médicales

Les scientifiques vont maintenant tester leur méthode en milieu clinique, où demeure une certaine appréhension vis-à-vis des IA. «Construire la confiance sur l'évaluation des IA est une étape clé vers leur adoption en milieu hospitalier», précise la Dre Mina Bjelogrlic, qui dirige l'équipe *Machine Learning* dans le service du Prof Lovis et deuxième auteure de cet article. «Notre étude porte sur l'évaluation des IA basées sur des séries temporelles, mais la méthodologie pourrait également s'appliquer à des IA portant sur d'autres modalités utilisées en médecine, comme l'image ou le texte.»

contact

Christian Lovis
Professeur
Directeur du Département de radiologie et informatique médicale
Faculté de médecine
UNIGE

Médecin-chef
Service des sciences de l'information médicale
HUG

+41 22 372 88 83

Christian.Lovis@unige.ch

Gianmarco Mengaldo
Professeur assistant
Directeur, MathEXLab
Dept of Mechanical Engineering
College of Design and Engineering
National University of Singapore (NUS)

+65 6516 8023

mpegim@nus.edu.sg

Hugues Turbé
Doctorant
Département de radiologie et informatique médicale
Faculté de médecine
UNIGE

+41 22 379 08 15

Hugues.Turbe@unige.ch

DOI: 10.1038/s42256-023-00620-w

UNIVERSITÉ DE GENÈVE
Service de communication
24 rue du Général-Dufour
CH-1211 Genève 4

Tél. +41 22 379 77 17

media@unige.ch
www.unige.ch