Avertissement: notes prises au vol… erreurs possibles… prudence!

HUG: Hôpital cantonal de Genève mardi 13 Juin 2023

L'interprétabilité des réseaux de neurones : créer la confiance

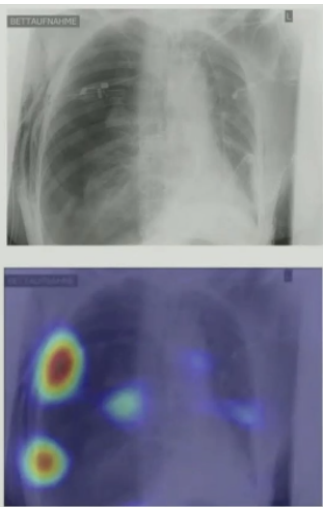
Mr Hugues Turbé, science de l'information médicale

Avant l'avènement des réseaux de neurones, les décisions étaient axées sur les connaissances, avec des variables mesurables et des règles claires, ce qui permettait de retourner sur celles-ci et de comprendre comment la décision a été prise.

Une décision prise par un réseau de neurones, est axée sur les données qui entraînent le réseau avec des milliards de paramètres… La décision qui en ressort, bien qu'ayant eu des résultats impressionnants, ne peut être vérifiée par un raisonnement.

En médecine en particulier, il y a un besoin de transparence à travers tous les acteurs impliqués: dans le développement, pour identifier les possibles biais du modèle et les erreurs systématiques, pour le côté légal, qui commence à voir le jour (AI act, FDA), et auprès des utilisateurs, pour prendre des décisions informées et donner confiance.

La transparence a plusieurs aspects… annoncer que la décision est influencée par un modèle d'intelligence artificielle (IA), expliquer ou comprendre comment le modèle fonctionne, et comprendre comment et quelles variables fournies influencent le résultat (Post-hoc interpretability).



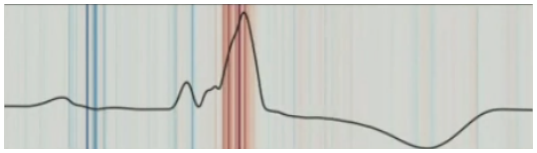
Le travail présenté aujourd'hui se concentre surtout sur ce dernier point.

Un modèle a été entraîné sur des radiographies, pour détecter un pneumothorax. L'IA est ensuite testée sur les radios encore jamais vues, et donne de très bons résultats.

En regardant quelle variable l'influence (en traçant le regard de l'IA), c'est en fait la présence du drain qui lui met la puce à l'oreille… et l'IA ne détecte aucun pneumothorax en l'absence d'un drain, elle ne regarde donc pas le parenchyme pulmonaire…

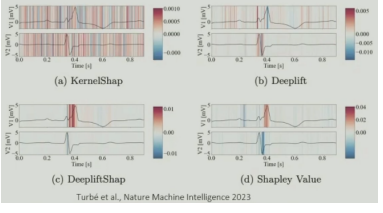
Cela montre qu'au-delà de la performance, il faut se concentrer sur la compétence d'un modèle. Et pour cela, l'interprétabilité est essentielle: il faut savoir quelles sont les variables qui influencent le résultat.

L'exemple qui suit est basé sur une image d'ECG. Un "relevancy score" évalue la contribution de chaque pixel sur la prédiction du modèle et donne une carte de contribution.



Rouge: en faveur
Bleu: en défaveur

Cependant, il existe de nombreuses méthodes d'interprétabilité, et appliquées sur les mêmes données et le même modèle, elles donnent des résultats différents.



L'équipe a donc décidé d'évaluer quelle est la meilleure méthode pour interpréter les données utilisées par l'IA.

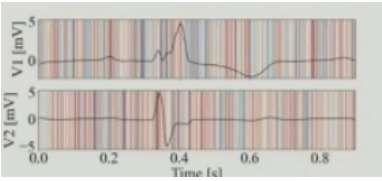
Trois biais d'évaluation de ces méthodes sont observés:

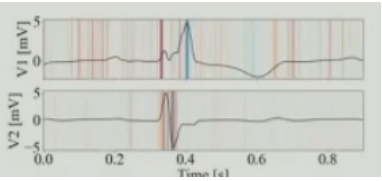
- influence du jugement humain: parce que la méthode "relevancy score" montre une utilisation des données similaires au raisonnement humain, il y a une tendance à reconnaître celle-ci comme la méthode juste.

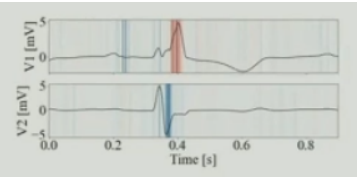
- Changement de distribution: lors d'évaluation des méthodes d'interprétabilité, la distribution des variables d'entrée (les variables qui ont entraîné le modèle) est modifiée, ce qui biaise l'évaluation.

- Ré-entraînement: pour éviter les changements de distribution, le modèle doit être ré-entraîné, ce qui mène à l'évaluation d'un modèle annexe à celui qui devait être évalué…

Le groupe a développé une méthode d'évaluation des méthodes d'interprétabilité, ce qui leur permet de dire que la carte de gauche est bruitée, celle du milieu est approximative et celle de droite est précise et donc correcte…







Après tant de précision… je peine à comprendre pourquoi la méthode n'est pas expliquée un minimum…ce travail a fait l'objet d'une publication citée plus bas, pour les personnes versées.

Conclusions

- Il est nécessaire d'améliorer le niveau de preuve pour les réseaux d'IA
    - Évaluer la pertinence de l'IA pour répondre à une question donnée (exemple du pneumothorax)
    - Éliminer les possibles biais, notamment dans les méthodes d'interprétabilité
- L'évaluation des méthodes d'interprétabilité doit être non-biaisée, autant techniquement que par le jugement humain
- L'adoption en clinique de l'IA dépendra de la confiance en ces modèles

Pour plus de détails, l'article récemment publiée par Turbé et al: Evaluation of post-hoc interpretability methods int time-series classification

Questions

Q: Est-il possible d'appliquer cette recherche de d'interprétation de la réponse chez des IA génératives comme chat GPT?
R: Oui, il est possible d'identifier quels mots d'entrée sont responsables de la réponse reçue.

Commentaire: Le big data n’est pas le pluriel de l’anecdote… La meilleure réponse n’est pas forcément la plus fréquente… et donc une limite de l’IA est de donner une réponse singulière.

Compte-rendu de Valentine Borcic
valentine.borcic@gmail.com
Transmis par le laboratoire MGD
colloque@labomgd.ch