Proposition de stage de recherche M2

« Explicabilité des réseaux profonds au moyen d'approches symboliques »

Laboratoire ICube de Strasbourg, équipes SDC et CSTB

Durée : 5 à 6 mois

Encadrement : Stella Marc-Zwecker (stella@unistra.fr), Anne Jeannin-Girardon (jeanningirardon@unistra.fr), Delphine Bernhard (dbernhard@unistra.fr)

Contexte

Les systèmes autonomes intelligents dits à "boîte noire", qui reposent sur des algorithmes d'apprentissage comme les réseaux de neurones profonds, deviennent omniprésents dans notre paysage quotidien. L'exigence d'un meilleur encadrement de ces algorithmes devient un enjeu sociétal, et requiert le développement de techniques permettant de comprendre leur fonctionnement ou d'expliquer leurs décisions. Ce stage s'inscrit dans le cadre du projet DEEPISH (Deep lEarning ExPlainabilIty through Symbolic approacHes), qui a pour objectif de proposer un modèle reposant sur des techniques de raisonnement symbolique (graphes de connaissances et règles), permettant d'expliquer les décisions de systèmes basés sur un apprentissage profond. Le domaine d'application considéré est le diagnostic médical.

Différentes approches ont été proposées pour développer l'explicabilité des modèles profonds. Parmi les plus populaires, on trouve les techniques de visualisation permettant d'identifier, dans le cadre de la reconnaissance d'objets dans des images, les portions de celles-ci ayant permis au modèle de faire sa prédiction (Wang et al., 2020). Bien qu'intéressants, ces modèles sont limités lorsque les objets recherchés sur l'image sont trop complexes pour être appréciés à l'œil nu par des experts humains, par exemple dans le cas de détection de lésions subtiles dans des mammographies (Oren et al., 2020).

Nous envisageons une approche multi-modale qui permettrait d'identifier les facteurs de confusion dans les données. En effet, dans le domaine médical, de nombreuses sources de données peuvent apporter des éléments permettant d'appuyer ou de rejeter un diagnostic : rapports textuels, bilans sanguins, données génétiques, etc. On peut alors concevoir un système, qui, lorsqu'une lésion non détectable par un expert humain est caractérisée, pourrait fournir d'autres éléments factuels appuyant sa prédiction : ***si*** *le patient est une femme* ***et*** *que la patiente possède le marqueur génétique xxx* ***alors*** *il est probable à n% qu'un traitement soit nécessaire*.

Objectif du stage

Il sera d'abord nécessaire, pour alimenter le système de raisonnement, de commencer par construire des graphes de connaissances à partir de données textuelles issues des données multi-modales (coupes histologiques et rapports histologiques) disponibles, afin d'en extraire des concepts qui seront utilisés par le système de raisonnement. L'extraction d'informations à partir de textes nécessite d'extraire des triplets comprenant un sujet, une relation et un objet

(Hohenecker et al., 2020, Solawetz & Larson, 2021). Ces graphes de connaissances seront ensuite enrichis par des connaissances extraites automatiquement à partir d'articles scientifiques disponibles dans le domaine public.

On pourra ensuite considérer que le modèle profond utilisé n'aura pas d'autre utilité que d'extraire des faits à partir de données complexes (ce qu'aucun système à base de règle n'est capable de faire), qui viendront compléter la connaissance organisée extraite des données textuelles. Ainsi, dans un deuxième temps, il faudra étudier différents types d'approches permettant de générer des règles logiques de façon autonome, comme les approches neuro-symboliques (Garcez et al., 2019 ; Ciravegna et al., 2021), ou les systèmes de classeurs (*Learning Classifier Systems*) (Orhand et al., 2021).

Pré-requis : l'étudiant·e en M2 informatique ou de niveau équivalent, devra avoir une spécialisation en intelligence artificielle ("deep learning", modélisation de connaissances, raisonnement logique). Il ou elle devra maîtriser le langage Python, être capable de manipuler les bibliothèques de TAL (spaCy, stanza, flair) et les réseaux de neurones, et être autonome pour l'implémentation.

Bibliographie

Ciravegna, G., Barbiero, P., Giannini, F., Gori, M., Lió, P., Maggini, M., & Melacci, S. (2021). Logic Explained Networks. *arXiv preprint arXiv:2108.05149*.

Garcez, A. D. A., Gori, M., Lamb, L. C., Serafini, L., Spranger, M., & Tran, S. N. (2019). Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning. *arXiv preprint arXiv:1905.06088*.

Hohenecker, P., Mtumbuka, F., Kocijan, V., & Lukasiewicz, T. (2020). Systematic Comparison of Neural Architectures and Training Approaches for Open Information Extraction. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 8554-8565.

Oren, O, Gersh, B. J. and Bhatt, D. L., "Artificial intelligence in medical imaging: switching from radiographic pathological data to clinically meaningful endpoints," The Lancet Digital Health, vol. 2, no. 9, pp. e486–e488, Sep. 2020, doi: 10.1016/S2589-7500(20)30160-6.

Orhand, R., Jeannin-Girardon, A., Parrend, P. and, Collet, P., "Explainability and Performance of Anticipatory Learning Classifier Systems in Non-Deterministic Environments", Genetic and Evolutionary Computation Conference (GECCO), Lille, France, juillet 2021

Solawetz, J., & Larson, S. (2021). LSOIE : A Large-Scale Dataset for Supervised Open Information Extraction. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2595-2600.

Wang et al., "Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2020

Zhang, Q., Cao, R., Shi, F., Wu, Y. N., & Zhu, S. C. (2018, April). Interpreting cnn knowledge via an explanatory graph. In *Thirty-Second AAAI Conference on Artificial Intelligence*.