Graduate Research School
University of Bordeaux
LaBRI (C.S. Lab in Bordeaux)

Stage de Master Recherche
(Intelligence Artificielle)

I.A. de Confiance :
Encadrement de Décisions issues
De l'Apprentissage Automatique
par des Preuves Formelles

Démarrage Fév. 2023
(Un financement de thèse est associé à ce stage)

Mises à jour: https://www.labri.fr/perso/lsimon/fr/research/internships

---

Encadrement
Laurent Simon, Professor at the Enseirb Matmeca / LaBRI (Bordeaux)
Head of the Computer Science teaching department of the ENSEIRB-Matmeca engineering school
Head of the industrial chair "Trustworthy AI" of Fondation Bordeaux Université
Co-Head of the A.I. axis of the LaBRI
Chair of the French Association on Constraint Programming

lsimon@labri.fr

Thèmes de recherche en lien avec l'IA.
Machine Learning, Reinforcement Learning, SAT, Logic, Constraint Programming, Trustworthy AI

Localisation
LaBRI, Laboratoire Bordelais de Recherche en Informatique, Talence, France

(This PhD will be fully funded by the chair "towards a trustworthy A.I." led by Laurent Simon and supported by the Fondation Bordeaux Université)

Description

Dans le but d'atteindre la meilleure précision possible, les systèmes de décisions/recommandations construits à l'aide d'apprentissage automatique agissent principalement sur un levier : la complexité calculatoire des fonctions apprises (*e.g.* le nombre de paramètres appris). Il est ainsi d'usage de bâtir des systèmes de recommandations mettant en jeu des millions de calculs conduisant à une décision, ce qui a pour effet immédiat de rendre, par construction, l'inspection de leur comportement impossible en termes simples et compréhensibles. Viser uniquement la précision peut avoir d'immenses intérêts applicatifs mais dès lors que leur mise en œuvre requiert de **comprendre**, **justifier**, **expliquer**, ou **garantir** certaines décisions, on voit que seule la précision ne suffit pas.

Comment est-il possible d'avoir confiance dans une décision dont le calcul est par construction impossible à résumer en termes compréhensibles ?

Ce constat a entraîné la multiplication de travaux d'un nouveau sous-domaine de l'intelligence artificielle, en lien avec la problématique de **confiance** dans les décisions autonomes (ainsi, de nombreux *workshops* spécialisés sont proposés en marge de toutes les plus grandes conférences en IA).

Suivant les applications visées, la confiance peut revêtir différentes formes. Dans l'approche que nous proposons, la confiance est exprimée en termes de **garanties prouvées**. Pour cela, nous nous appuierons sur les progrès réalisés ces dernières années par **les méthodes formelles** (notamment grâce à l'emploi de solveurs SMT / SAT) de manière à encadrer des systèmes décisionnels "boîte noire" par des **systèmes plus simples permettant d'exprimer les garanties visées** dans des langages accessibles. Si les méthodes formelles ont également fait d'immenses progrès ces dernières années, tenter de les utiliser directement sur les systèmes ayant des milliards de paramètres semble vain. Nous proposons dans cette approche d'encadrer ces systèmes par d'autres systèmes plus simples, et de garantir les propriétés sur les fonctions encadrant ces systèmes.

Ainsi, étant donné par exemple un réseau de neurones *RNNA* prédisant "Oui" ou "Non" sur une entrée quelconque, nous proposons de construire deux formules logiques *Up* et *Down* permettant d'encadrer au mieux les décisions de RNNA (*Up* pour les "Oui", *Down* pour les "non") et offrant des garanties **d'explicabilité** ou permettant l**a preuve de certaines propriétés désirées**. Le point clé est que les langages Up et Down sont définis sur des langages logiques permettant la preuve formelle (logiques propositionnelles, logiques temporelles, …) tout en capturant un maximum de précision depuis *RNNA*.

Cette approche pose de nombreux problèmes. La difficulté principale consiste à construire *Up* et *Down* à partir de A. Il conviendra donc d'étudier comment le langage avec lequel *RNNA*, *Up* et Down sont exprimés permettra de réécrire A en minimisant les pertes de précision des fonctions *Up* et *Down*. Des approches issues des **Réseaux de Neurones Binarisés**, ainsi que les approches à base de **Compilation de base de connaissances** pourront être étudiées. Suivant les connaissances du stagiaire, il pourra également être étudié **comment l'apprentissage en lui-même peut être modifié pour viser directement l'apprentissage des fonctions *RNNA*, *Up* et *Down*** plutôt que de les calculer a posteriori.

Suite du stage

Vu la richesse du sujet, le stage est associé à **une bourse de 3 ans de thèses,** co-financée par la Chaire IA Digne de Confiance et le projet RobSYS.

Candidat

Le candidat devra avoir un excellent niveau scientifique et une bonne connaissance des méthodes de raisonnement basés sur la logique. De plus, il est également attendu un très bon niveau de programmation.

Le stage pourra commencer dès février 2023.

Contact : Laurent Simon lsimon@labri.fr