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

Master : Research Internship in Artificial Intelligence

Trustworthy AI: Bounding Machine Learning Decisions by Proven Systems

Starting Feb. 2023
(A PhD Grant is associated with this internship)

Updates: https://www.labri.fr/perso/lsimon/research/internships

---

Supervisor
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

Disciplinary fields related to AI
Machine Learning, Reinforcement Learning, SAT, Logic, Constraint Programming, Trustworthy AI

Localization
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

In order to achieve the best possible accuracy, decision/recommendation systems built with the help of machine learning act mainly on one lever: the computational complexity of the learned functions (e.g. the number of learned parameters). It is thus common practice to build recommendation systems involving millions of calculations leading to a decision, which has the immediate effect of making it impossible, by construction, to inspect their decision in simple and understandable terms. Aiming solely at precision may have immense applicative interests, but as soon as the implementation requires **understanding**, **justifying**, **explaining** or **guaranteeing** certain decisions, we see that precision alone is not enough when trust is needed.

How is it possible to have confidence in a decision whose calculation is by construction impossible to summarize in comprehensible terms?

This observation has led to the multiplication of work in a new subfield of artificial intelligence, linked to the problem of **trust in autonomous decisions** (thus, many specialized workshops are proposed in the margins of all the major AI conferences).

Depending on the application, trust can take different forms. In the approach we propose, **trust is expressed in terms of proven guarantees on the final system**. For this purpose, we will rely on the progress made in recent years by formal methods (in particular thanks to the use of SMT / SAT solvers) so as to bound "black box" decision systems by simpler systems allowing to express the targeted guarantees in accessible languages. Although formal methods have also made immense progress in recent years, attempting to use them directly on systems with billions of parameters seems hopeless. In this approach, we propose to bound these systems by other, simpler, systems, and to guarantee the properties on the functions bounding these systems.

Thus, given for example *RNNA* a Neural network predicting "Yes" or "No" on any input, we propose to build two logical formulas *Up* and *Down* allowing to bound RNNA's decisions as precisely as possible (*Up* for "Yes", *Down* for "No") and offering guarantees of explainability or allowing the proof of some desired properties. The key point is that *Up* and *Down* languages will be defined on a formal language allowing formal proof (propositional logics, temporal logics, ...) while capturing a maximum of precision from *RNNA*.

The main difficulty in the internship will probably be constructing *Up* and *Down* from *RNNA*. It will therefore be necessary to study how the language in which *RNNA*, *Up* and *Down* are expressed will allow A to be rewritten while minimizing the loss of precision of the *Up* and *Down* functions. Approaches based on Binarized Neural Networks, as well as approaches based on Knowledge Base Compilation can be studied. Depending on the candidate's knowledge, it may also be possible to study how the learning itself can be modified to directly target the learning of *RNNA*, Up and *Down* functions together rather than computing them *a posteriori*.

Continuation of the internship

Given the richness of the subject, the internship is associated with a 3-year thesis grant, co-financed by the Trustworthy AI Chair and the RobSYS project.

Candidate

The candidate should have an excellent scientific background and a good knowledge of logic-based reasoning methods. In addition, a very good level of programming is also expected.

The internship can start as early as February 2023.

Contact: Laurent Simon lsimon@labri.fr