Offre n°2025-08697

PhD Position F/M Machine Learning Trustability : Learning and Verification of Soft Automata

Type de contrat : Fixed-term contract

Niveau de diplôme exigé : Graduate degree or equivalent

Fonction : PhD Position

A propos du centre ou de la direction fonctionnelle

The Inria Centre at Rennes University is one of Inria's nine centres and has more than thirty research teams. The Inria Centre is a major and recognized player in the field of digital sciences. It is at the heart of a rich R&D and innovation ecosystem: highly innovative PMEs, large industrial groups, competitiveness clusters, research and higher education players, laboratories of excellence, technological research institute, etc.

Contexte et atouts du poste

The candidate will be part of the collaborative project **SAIF, "Safe AI through Formal methods,"** (https://project.inria.fr/saif/), that involves renowned research labs in CS : Inria, CEA-List, LIX, LaBRI, LMF.

Mission confiée

Context

Reconstructing a dynamic system from its trajectories is an old topic, addressed by several communities. This is called system identification in systems theory, equation discovery in physics, and automata learning in computer science (CS). In CS, one may wish to recover an automaton from words of its language and possibly from counter-examples. Classical "exact" algorithms exist to do so, as the celebrated L-star, but they rely on powerful oracles, i.e. on the possibility to make queries to the unknown system. Modern machine learning techniques now provide an alternative approach, through various neural network (NN) architectures. Beyond their impressive performances, they also enjoy appealing features :

- They are passive methods, relying simply on data-bases of examples (no queries, no need for powerful oracles, even no need for counter-examples).
- They generalize extremely well and can be used as generators.
- Focusing on the architectures behind Large Language Models (LLM), they manage to capture global features that go beyond classical regularity (spelling, grammar, syntax) as for example style or even meaning.

Important down sides remain, however : these new models are huge, very different from the traditional formalisms handled by formal methods, their behavior is poorly understood... while one would like to assess their safety and reliability for numerous critical applications.

Objectives

The objective of this PhD is to explore the way various NN-based architectures manage to approximate formal languages, i.e. learn surrogate automata from their traces. Beyond well established results on the expressive power of these models, the focus will be on the capabilities of the pair model + learning algorithm. Several authors have shown that almost discrete behaviors emerge naturally when NN are trained by automata traces, despite their definition as continuous state space systems, whence the name "soft automata." Another objective will be to assess the robustness and reliability of such NN-based models as automata approximators, by means of appropriate formal methods.

Several research directions are envisioned, that will be adapted to the skills and wishes of the candidate. We mention some of them below.

- Exploring the approximation abilities of recurrent neural networks (RNN). RNNs are good approximators of regular languages, but tend to build quasi discrete approximations resembling local automata or n-gram models. This property has to be further understood by examining how well RNN learn more complex languages, and by measuring the distance between the original language and the one approximated by the RNN. This is both an experimental and a theoretical direction, as no algorithms yet exist to estimate such distances.
- Exploring the robustness of the models learnt by RNN, to identify stable regions of their state space and unstable ones. The effect of extra data,

missing data or poisoned data on such robustness also has to be characterized. This research track will also aim at learning more robust models, by enforcing properties of the hidden state space, or by enforcing specific safety properties.
- Replacing a true automaton by its RNN surrogate (used as a generative model for example) raises questions like its reliability. One would like to verify properties of runs produced by such soft automata, for example safety properties. Few algorithms yet exist for model checking such models, and they mostly focus on static feed-forward NN, not recurrent ones.
- Exploring the properties of other architectures. While RNN have a vanishing memory, other structures like GRU or LSTM provide longer term memory, not to mention Transformers or attention-based architectures. The approximation abilities of such models have to be better understood, in particular to characterize the family of languages they best suit. New NN architectures and learning algorithms will be explored, with the aim to better capture multiresolution features of a run that best predict the future of this run. For example to better learn hierarchical automata.

Related bibliography

- Dana Angluin's L* algorithm, "Learning Regular Sets from Queries and Counter-Examples," 1987.
- Frits Vaandrager, Bharat Garhewal, Jurriaan Rot, Thorsten Wissmann : "A New Approach for Active Automata Learning Based on Apartness," 2022.
- Gail Weiss, Yoav Goldberg, Eran Yahav : "On the Practical Computational Power of Finite Precision RNNs for Language Recognition," 2018.
- Ilya Sutskever, James Martens, Geo?rey Hinton : "Generating Text with Recurrent Neural Networks," ICML 2011
- J. Michalenko, A. Shah, A. Verma, R. Baraniuk, S. Chaudhuri, A. Patel : "Representing Formal Languages : A Comparison Between Finite Automata and Recurrent Neural Networks," ICLR 2019.
- Zeyuan Allen-Zhu, Yuanzhi Li, "Physics of Language Models : Part 1, Learning Hierarchical Language Structures," 2023, ICML 2024 tutorial.

Avantages

- Subsidized meals
- Partial reimbursement of public transport costs
- Possibility of teleworking (90 days per year) and flexible organization of working hours
- Partial payment of insurance costs

Rémunération

2 200 euros per month

Informations générales

- **Ville :** Rennes
- **Centre Inria :** Centre Inria de l'Université de Rennes
- **Date de prise de fonction souhaitée :** 2025-09-01
- **Durée de contrat :** 3 years
- **Date limite pour postuler :** 2025-05-31

Contacts

- **Équipe Inria :** AT-REN
- **Directeur de thèse :**
  Fabre Eric / eric.fabre@inria.fr

A propos d'Inria

Inria est l'institut national de recherche dédié aux sciences et technologies du numérique. Il emploie 2600 personnes. Ses 215 équipes-projets agiles, en général communes avec des partenaires académiques, impliquent plus de 3900 scientifiques pour relever les défis du numérique, souvent à l'interface d'autres disciplines. L'institut fait appel à de nombreux talents dans plus d'une quarantaine de métiers différents. 900 personnels d'appui à la recherche et à l'innovation contribuent à faire émerger et grandir des projets scientifiques ou entrepreneuriaux qui impactent le monde. Inria travaille avec de nombreuses entreprises et a accompagné la création de plus de 200 start-up. L'institut s'e?orce ainsi de répondre aux enjeux de la transformation numérique de la science, de la société et de l'économie.

L'essentiel pour réussir

The ideal candidate should have a taste for formal methods and abilities for experimental work using standard machine learning libraries. Skills in statistics, data science, or personal projects in machine learning are a plus.

Position open to candidates with an MsC in Computer Science (orientation to formal aspects of CS preferred), or with an Engineering degree in CS.

Attention: Les candidatures doivent être déposées en ligne sur le site Inria. Le traitement des candidatures adressées par d'autres canaux n'est pas garanti.

Consignes pour postuler

Please submit online : your resume, cover letter and letters of recommendation eventually

Sécurité défense :
Ce poste est susceptible d'être affecté dans une zone à régime restrictif (ZRR), telle que définie dans le décret n°2011-1425 relatif à la protection du potentiel scientifique et technique de la nation (PPST). L'autorisation d'accès à une zone est délivrée par le chef d'établissement, après avis ministériel favorable, tel que défini dans l'arrêté du 03 juillet 2012, relatif à la PPST. Un avis ministériel défavorable pour un poste affecté dans une ZRR aurait pour conséquence l'annulation du recrutement.

Politique de recrutement :
Dans le cadre de sa politique diversité, tous les postes Inria sont accessibles aux personnes en situation de handicap.