Machine Learning Trustability : Verification of Soft Automata

PhD / MsC internship

INRIA Rennes - France (Brittany)

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 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 for numerous critical applications.

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." The focus will be on assessing the robustness and reliability of such NN-based models as automata approximators.

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.

This PhD proposal can be downsized to an MsC internship. The ideal candidate should have a taste for formal methods and abilities for experimental work using standard machine learning libraries.

The PhD will take place at INRIA Rennes (Brittany, France). 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.

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, Geoffrey 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.

Contact : Eric Fabre, eric.fabre@inria.fr, +33 (0)2 99 84 73 26