Trustable Machine Learning :
Analysis and Verification of Soft Automata

PhD position

INRIA Rennes - France (Brittany)

Learning automata from their traces has long been addressed from a purely logical perspective (e.g. Angluin's L* algorithm), until neural architectures offered an amazing alternative : ground breaking performances, summoning models at the boundary between the continuous world and the discrete world, leveraging probabilistic approaches... but providing no guarantees on the models produced by the learning algorithms ! The objective of this thesis is to shed light on the properties of these "soft automata," based on neural networks, by crossing perspectives from system theory, statistics, optimization and formal methods in order to provide guarantees on these dynamic systems, to understand their expressivity, their robustness to noise and attacks, and their sensitivity to data quality. The thesis will examine different architectures, from plain recurrent neural networks to gating and attention mechanisms, and up to more recent architectures like state space models or Mamba. The design of new neural architectures with better properties, and the design of jailbreaking and poisoning attacks to these models are also in the scope. More details below.

The adaptation of verification techniques to neural networks (NN) has (successfully) focused on a rather narrow topic : how robust is the output of a NN to perturbations on the input. Standard approaches are borrowed to static analysis, and perform reasonings at the scale of individul neurons. Besides scalability issues, these methods are oriented to classifiers and hardly adapt to models of dynamic systems. Mostly, they put aside the huge engineering effort that led to high performance neural architectures. This is the angle adopted here : exploiting this architecture to tailor verification approaches.

Numerous neural architectures have been designed to identify dynamic systems from their traces. We focus here on the learning of automata from part of their language. These models are trained as predictors of the future, from positive examples only, and not as classifiers (deciding if some imput word is in the language or not). This makes them generative models, that could be used as surrogate of automata, whence the generic name of "soft automata" as these models compute in $\mathbb{R}^d$.

Recurrent neural networks (RNN) are the most natural neural architecture that comes to mind when one wants to learn an automaton. While trained with gradient descent, these objects have been shown to converge to discrete behaviors : their state space tends to form clusters which structure and properties are still under investigation. Similar behaviors appear with variants like LSTM or GRU, that inrtroduce gating mechanisms in order to prevent the fast memory decay of plain RNN. These emerging properties suggest that understanding the structuration of the state space of these models is key to address questions like their robustness to noise, to data quality and to attacks.

Independently, the success of transformers in text modeling/generation has motivated their adaptation to the larger domain of time series analysis. It is yet unclear if foundation models could emerge in that field, but successful attempts have been reported with rather simple architectures. The simplest is probably PatchTST, which abilities to learn automata remain to be explored (taking words in the language as time series). A possible research direction could be to identify how the attention mechanism and the sketching of patches in a time series combine to identify features in a sequence, and further to structure the state space of these models. Still with the aim of assessing their generalization abilities and their robustness to noise or attacks.

More recently, other architectures have been introduced under the generic term of "state space models," like HiPPO or S4, and further Mamba. While originally addressing two limitations of transformers, a finite window context and a quadratic computational cost in the size of this window, they take inspiration from well known linear models in systems theory, and open the way to a more interpretable state space. A possible direction of the thesis could therefore be to explore the relevance of these models as surrogate automata, and again make use of their internal structure to design analysis and verification techniques.

The 3 research directions mentioned above will not all be explored at the same level. The topic will be adapted to the candidate. The ideal candidate should have a solid background in mathematics, 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 Computet Science : Inria, CEA-List, LIX, LaBRI, LMF, ENS Paris, ENS Saclay.

Bibliography :

- Gail Weiss, Yoav Goldberg, Eran Yahav : "On the Practical Computational Power of Finite Precision RNNs for Language Recognition," 2018.
- 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.
- Albert Gu, Tri Dao, "Mamba : Linear-Time Sequence Modeling with Selective State Spaces," 2024, https ://doi.org/10.48550/arXiv.2312.00752
- Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, Jayant Kalagnanam : "A time series is worth 64 words : long-term forecasting with transformers," ICLR 2023.

Contact : Eric Fabre, eric.fabre@inria.fr

To apply : `https://recrutement.inria.fr/public/classic/fr/offres/2026-10158`