Automata learning from Recurrent Networks via Clustering

Keywords: Clustering, Automata Learning, Recurrent Neural Network, Interpretability

Institution

The French Alternative Energies and Atomic Energy Commission (CEA) is a key player in research, development, and innovation. Drawing on the widely acknowledged expertise gained by its 16,000 staff spanned over 9 research centers with a budget of 4.1 billion Euros, CEA actively participates in more than 400 European collaborative projects with numerous academic (notably as a member of Paris-Saclay University) and industrial partners. Within the CEA Technological Research Division, the CEA List institute addresses the challenges coming from smart digital systems.

Among other activities, CEA List's Software Safety and Security Laboratory (LSL) research teams design and implement automated analysis in order to make software systems more trustworthy, to exhaustively detect their vulnerabilities, to guarantee conformity to their specifications, and to accelerate their certification. The lab recently extended its activities on the topic of AI trustworthiness and gave birth to a new research group: AISER (Artificial Intelligence Safety, Explainability and Robustness).

Scientific context

Through the recent developments of AI, their use has become even more widespread, even in industrial settings. Nevertheless, studies are flourishing showing the dangers that such AI can bring, whether in terms of safety, privacy or fairness. We can for example cite the case of adversarial attacks, where small perturbations invisible to naked eyes, can drastically change the output of an AI model. To face these dangers, works and tools are constantly emerging to increase the trust one can have in AI systems.

One of the tool developed at CEA in the AISER team is PyRAT, a Python tool based on abstract interpretation techniques to assess the robustness of a neural network in face of perturbations. Works to extend PyRAT for different neural networks architectures and their properties verification are undergoing. For Recurrent Neural Networks (RNNs), the task of verifying its properties by abstract interpretation is quite challenging due to its recurrent nature. Learning an automata as a faithfull surrogate model to the original RNN is an approach that enables us to reduce its complexity, thus easing the verification step. Passive automata learning from RNNs is an approach largely based on clustering techniques. The idea is to group RNNs hidden states, which carry similar semantics, into abstract states, then draw abstract transitions between those abstract states by following the input trajectories in the hidden states space.

Internship

The subject of this internship adresses the predominant reliance on k-means clustering in passive automata learning from RNNs execution traces. K-means operates under the assumption of isotropic (spherical) cluster variance and struggles with complex, high dimensional, and real world data, which will often impact the size (number of states) and fidelity (faithfulness to the original RNN behavior) of the extracted automata.

The main objectives of this internship is to investigate, implement, and evaluate clustering algorithms for extracting automata representations from trained recurrent neural networks, starting with the following steps:

- First the intern will be guided to study and get familar with passive automata learning from RNNs as well as SoTA clustering algorithms to start with (a comparative study will be made)
- Next, algorithms implementation and comparison on pre-defined benchmarks (textual data, tabular data, etc.)
- If time permits, multiple axes can be tackled during this internship:
  - Improving or proposing a novel clustering-based automata learning algorithm
  - Working on interpretability and explainability of Recurrent Neural Networks.

This work will have contributions to the field of automata learning and to neural networks interpretability. The internship will very likely conclude by publishing a paper (workshop, conference) depending on the quality of the work to be carried.

Qualifications

The candidate will work at the crossroads of formal methods and machine learning. As it is not realistic to be expert in both fields, we encourage candidates that do not meet the full qualification requirements to apply nonetheless. We strive to provide an inclusive and enjoyable workplace. We are aware of discriminations based on gender (especially prevalent on our fields), race or disability, we are doing our best to fight them.

- **Minimal**
  - Master student or equivalent (2nd/3rd engineering school year) in computer science
  - knowledge of Python
  - ability to work in a team, some knowledge of version control
- **Preferred**
  - notions of clustering algorithms and neural networks
  - notions of automata theory or graph theory

Characteristics

The candidate will be monitored by two research engineers, and one PhD student of the team.

- **Duration:** 4 to 6 months from early 2026
- **Location:** CEA Nano-INNOV, Paris-Saclay Campus, France
- **Compensation:**
  - €1300 if you are in M1/second year of engineering school, €1400 if you are in M2/third year of engineering school
  - maximum €229 housing and travel expense monthly allowance (in case a relocation is needed)
  - CEA buses in Paris region and 75% refund of transit pass
  - subsidized lunches
  - 2 days of remote work

Application

If you are interested in this internship, please send to the contact persons an application containing:

- your resume;
- a cover letter indicating how your curriculum and experience match the qualifications expected and how you would plan to contribute to the project;
- your bachelor and master 1 transcripts;

Applications are welcomed until the position is filled. Please note that the administrative processing may take up to 3 months.

Contact persons

For further information or details about the internship before applying, please contact:

- Augustin Lemesle (augustin.lemesle@cea.fr)
- Tristan Le Gall (tristan.le-gall@cea.fr)
- Jaouhar Slimi (jaouhar.slimi@cea.fr)

References

Hong, D., Segre, A.M., & Wang, T. (2022). AdaAX: Explaining Recurrent Neural Networks by Learning Automata with Adaptive States. Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.

Chen, X., & Güttel, S. (2022). Fast and explainable clustering based on sorting. Pattern Recognit., 150, 110298.