COLLEGE ED MATHEMATIQUES
DOCTORAL STIC
PAYS DE LA LOIRE

Nantes Université

THESE DE DOCTORAT

NANTES UNIVERSITE

ECOLE DOCTORALE N° 641
MAthématiques et Sciences et Technologies du numérique, de l'Information et de la Communication
SPécialité : Informatique et Architectures numériques

Par
Mohamed Reda Marzouk

Intelligibilité des réseaux de neurones récurrents par des machines à états finis

Thèse présentée et soutenue à Nantes, le 17 Octobre 2024
Unité de recherche : UMR 6004

Rapporteurs avant soutenance :

Nathanaël Fijalkow	Chargé de recherches, CNRS, Bordeaux
Céline Hudelot	Professeure, CentraleSupélec, Université Paris-Saclay, Paris

Composition du Jury :
Attention, en cas d'absence d'un des membres du Jury le jour de la soutenance, la composition du jury doit être revue pour s'assurer qu'elle est conforme et devra être répercutée sur la couverture de thèse

Président :	Prénom Nom	Fonction et établissement d'exercice (8)*(à préciser après la soutenance)*
Examinateurs :	Elisa Fromont	Professeure, Université de Rennes, Rennes
	Céline Hudelot	Professeure, CentraleSupélec, Université Paris-Saclay, Paris
	Nathanaël Fijalkow	Chargé de recherches, CNRS, Bordeaux
	Richard Dufour	Professeur, Université de Nantes, Nantes
	Rémi Eyraud	Maître de conférences, Université Jean Monnet, Saint-Étienne
Dir. de thèse :	Colin de La Higuera	Professeur, Université de Nantes, Nantes

COLLEGE ED MATHEMATIQUES
DOCTORAL STIC
PAYS DE LA LOIRE

Nantes Université

Titre : Intelligibilité des réseaux de neurones récurrents par interprétation avec des machines à états finis

Mots clés : ML interprétable, Réseaux de neurones récurrents, Machine à états finis

Résumé : Le paradigme de calcul neuronal, bien que performant, pose un défi majeur en termes d'analyse formelle, contrairement aux machines à états finis (FSM) qui offrent un paradigme plus formel et donc plus facilement analysable par des outils algorithmiques classiques avec des garanties théoriques strictes. Cette thèse explore les interactions et divergences entre ces deux paradigmes, en se focalisant sur le problème de l'interprétabilité des modèles et en mettant l'accent sur les aspects théoriques de ces connexions. En analysant la différence d'interprétabilité entre les réseaux de neurones récurrents (RNNs) et les automates pondérés (WAs) — une famille d'automates qui englobe à la fois les automates acceptant des langages binaires et les automates stochastiques implémentant des modèles de langage — sous l'angle de la théorie de la complexité computationnelle, nous montrons formellement que, contrairement aux RNN-ReLUs les automates pondérés (WAs) se prêtent plus facilement à une analyse interprétative à l'aide de métriques telles que le score SHAP et le score de polarisation. Ces résultats théoriques offrent un argument solide justifiant l'utilisation des modèles basés sur des automates comme abstraction des modèles neuronaux pour l'analyse de ces derniers. Dans un autre contexte, nous avons examiné la complexité associée à la quantification des disparités entre les modèles de langage basés sur des automates et les modèles de langage neuronaux. De nouveaux schémas d'approximation avec des garanties PAC ont été développés pour calculer l'entropie croisée entre les automates pondérés stochastiques et diverses familles de RNNs, notamment les LSTMs et GRUs.

Title : Intelligibility of Recurrent Neural Networks via Finite State Machines

Keywords : Interpretable ML, Recurrent Neural Networks, Finite State Machines

Abstract : Deep Learning has emerged as a dominant computing paradigm in the field of AI, demonstrating remarkable performance across a wide array of tasks. However, its inherent opacity presents significant challenges in terms of interpretability. Conversely, Finite State Machines (FSMs) represent a mature computing paradigm with established formal methods for analysis.

This thesis aims to explore intersections and divergences between these two paradigms by investigating their interpretability within the framework of computational complexity theory.

Our investigation uncovers a notable disparity in interpretability between RNNs and Weighted Automata (WAs). Specifically, we establish that while the complexity of conducting interpretative analysis on WAs, utilizing metrics such as the SHAP score and polarization score, is tractable, the same cannot be asserted for RNN-ReLUs. This observation highlights the potential of automata-based models as promising interpretation proxies for neural-based models, providing a compelling rationale for their incorporation in interpretative analyses.

In a separate context, we examined the complexity associated with quantifying disparities between automata-based and neural-based language models. To address this challenge, we developed novel approximation schemes with PAC guarantees for computing the cross-entropy between Stochastic Weighted Automata and various families of RNNs, including Long Short-Term LSTM/GRUs.