An introduction to the formal verification of neural networks

Tutorial proposal

Julien Girard-Satabin Sasha Cuau Guilhem Ardouin

Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France

julien.girard2@cea.fr sasha.cuau@cea.fr guilhem.ardouin@cea.fr

Short description

This tutorial proposes an introduction to the formal verification of neural networks, a field that aims to provide mathematical guarantees on the behaviour of neural networks. We will first motivates the need for such methods in the field of AI, followed by a presentation on the theoretical aspects of verification and a practical session.

Tutorial description

As machine learning systems become increasingly integrated into our everyday lives, ensuring their safety becomes even more crucial. Formal verification can provide guarantees in a provable way, making it a good candidate to assess whether a neural network behaves as expected.

This tutorial will introduce the application of formal verification techniques to neural networks [15]. It will first explore the reasons why formal methods are a crucial step towards AI systems' safety. Then, formal properties, such as robustness [5] and functional properties, will be introduced on an industrial example (ACAS-XU [13]). An emphasis on the different steps of the formal verification will be made: definition of the property on a system, translation to a specification language, and its effective verification. Verification techniques will be introduced, with a focus on abstract interpretation [14]. Then, state-of-the-art verification tools will be introduced, and the remaining part of the tutorial will be a practical session with one of them: PyRAT [12]. This tutorial will be concluded by open research questions regarding the efficient verification of properties and their formalisation.

Tutorial outline

1. Introduction: Motivation for the formal verification of neural networks

(This part will be supported by slides and interactive poll)

This part aims to introduce the topic of formal verification of neural network. An interactive session with a real life scenario of 10/15min will be held in order to give the intuition to the attendees:

- Real life scenario: an AI system for collision avoidance is embedded in a plane. The goal is to give them the intuition of safety and guarantees for critical system.
- Questions:
  - What could go wrong? (unexpected behaviour, perturbations…)
  - What would make you trust the system? (guarantees)
  - If time: Any ideas how to obtain those guarantees?
  - Why testing is not efficient?

There will be a quick presentation of the *ACAS-XU* benchmark [13] (Aircraft Collision Avoidance System), that will serve as a running example throughout the tutorial. This benchmark introduces properties regarding the expected behaviours of an aircraft collision avoidance system [10].

We will discuss the different potential weaknesses of neural networks and briefly give the intuition for the notions of safety. We will then motivate the need of formal methods to provide guarantees on the behaviour of neural networks.

2. Theoretical overview

(This part will be supported by slides)

This part will start with a formalisation of the verification problem, using first-order logic:

- introducing functional properties of ACAS-XU [10]
- introducing the local robustness property [5] (first with the intuition and then the mathematical definition)

Then, a bit of theory behind formal methods will be given:

- foundational concepts: Rice theorem, the notions of sound and complete algorithms
- different type of verification approaches (SMT, MIP, Abstract Interpretation)

- the theory of Abstract Interpretation (focus of the tutorial):
  - give the intuition for abstract interpretation, the technique of approximating the behaviour of a system as a compromise between precision and scalability
  - its application with neural network [17]
  - a simple example with interval arithmetic
  - introduction of a relational domain (zonotopes [16])
    - exact for linear operations
    - approximation for non-linear operations

3. Applications

(This part will be supported by a Jupyter notebook)

The application section will start with a step by step verification on a toy example and will be followed by a quick explanation on the functioning of provers:

- how to represent the models and properties?
- what are the different domains?
- what is the state-of-the-art of abstract interpretation-based tools? [12, 19]

A guided exercise will be given to the attendees to specify from scratch a specification on different models (including ACAS-XU) and to verify it with PyRAT (state-of-the-art abstract interpretation tool). In particular, the Jupyter notebook that will be provided to the attendees will contains for each example:

- a detailed description of the use case
- guided questions for the attendee to be able to formulate a property on the model
- guided questions for the attendee to write the property in VNN-LIB [8], the standard input language for neural network properties
- guided questions to launch the verification of the property using PyRAT

Potential openings

(This part will be supported by slides)

A non-exhaustive list of subjects that could be given as opening:

- Architecture challenges: verifying more complex architecture (i.e. transformers, RNNs)
- Using high level specification languages: avoiding writing pure VNN-LIB specifications (by using tools such as Vehicle [7] or CAISAR [1])

- Specification of more complex properties: fairness [4,11], confidence-based properties [3]...

By the end of the tutorial, the attendees should have:

- understood the basic concepts of formal verification of neural networks
- specified verification properties on neural network models
- verified these specifications by manipulating a dedicated prover

Targeted audience and expected knowledge

We will welcome people that are familiar with deep learning but with no specific training in formal methods. We expect attendees to be familiar with:

- Python basics
- NumPy and PyTorch

Objectives of this tutorial

This tutorial might interest people working in the AI field, notably because of its crucial importance for the integration of AI component in critical systems. In particular, this tutorial fits within the following objectives:

- Motiver et expliquer un sujet d'importance émergente pour l'IA.
- Introduire des experts (ou du moins des personnes familières avec l'IA) non spécialistes dans un sous-domaine de l'IA.

About the authors

Julien-Girard Satabin is a researcher at CEA LIST, Paris. They got their PhD on 2021 at Université Paris-Saclay, France. Their research interest are the topic of formal verification of machine learning programs, interpretable AI and symbolic reasoning. They authored several course on the topic of formal verification of AI and explainable AI, including a tutorial at European Conference on Artificial Intelligence 2024 and a 10h course at the European Summer School on Artificial Intelligence. They co-authored the following papers: [1,6,9,18]
Previous related work experiences:

- lectures given: https://julien.girard-satabin.fr/tutorials/ and https://julien.girard-satabin.fr/teaching/
- tutorial given at ECAI24: https://laiser.frama-c.com/laiser-websites/xai-ecai24/

- summer school lectures given at ESSAI25: https://caisar-platform.com/2025/06/30/essai.html

Guilhem Ardouin is a PhD Student at CEA List (French Atomic Energy Commission) and Paris-Saclay university. His research revolves around improving the efficiency of formal verification of neural networks, by proposing automated selection of algorithms to solve them, and on the expressivity of specification languages for the VNN problem [2]. He taught programming classes at CentraleSupélec and ENSTA Paris (C programming, Java OOP).
Website: https://gardouin.github.io/

Sasha Cuau is a PhD student at CEA List (French Atomic Energy Commission) and university Paris-Saclay. Her research focuses on the formal verification of transformers neural networks, including the approximation of the attention mechanism with abstract interpretation. She obtained a *best poster award* for "Towards the formal verification of the attention mechanism" at JITA 2025 (The Junior conference on Informatics: Theory and Applications). She gave a tutorial on the verification of neural network with PyRAT at CentraleSupélec.

References

[1] Michele Alberti, François Bobot, Julien Girard-Satabin, Alban Grastien, Aymeric Varasse, and Zakaria Chihani. The CAISAR Platform: Extending the Reach of Machine Learning Specification and Verification. In Ferruccio Damiani and Marie Farrell, editors, *Integrated Formal Methods*, volume 16194, pages 290–309. Springer Nature Switzerland.

[2] Guilhem Ardouin, Michele Alberti, and Julien Girard-Satabin. La confiance avec le contrôle : spécification et vérification d'hyperpropriétés sur réseaux de neurones. In *JFLA 2026 – 37es Journées Francophones des Langages Applicatifs*, volume JFLA 2026 – 37es Journées Francophones des Langages Applicatifs, Oberbronn, Alsace, France, January 2026. Marie Kerjean and Yannick Zakowski.

[3] Anagha Athavale, Ezio Bartocci, Maria Christakis, Matteo Maffei, Dejan Nickovic, and Georg Weissenbacher. Verifying Global Two-Safety Properties in Neural Networks with Confidence.

[4] Sumon Biswas and Hridesh Rajan. Fairify: Fairness Verification of Neural Networks.

[5] Marco Casadio, Ekaterina Komendantskaya, Matthew L. Daggitt, Wen Kokke, Guy Katz, Guy Amir, and Idan Refaeli. Neural Network Robustness as a Verification Property: A Principled Case Study.

[6] Lucas C. Cordeiro, Matthew L. Daggitt, Julien Girard-Satabin, Omri Isac, Taylor T. Johnson, Guy Katz, Ekaterina Komendantskaya, Augustin Lemesle, Edoardo

Manino, Artjoms Šinkarovs, and Haoze Wu. Neural Network Verification is a Programming Language Challenge.

[7] Matthew L. Daggitt, Wen Kokke, Robert Atkey, Ekaterina Komendantskaya, Natalia Slusarz, and Luca Arnaboldi. Vehicle: Bridging the Embedding Gap in the Verification of Neuro-Symbolic Programs.

[8] Stefano Demarchi, Dario Guidotti, Luca Pulina, and Armando Tacchella. Supporting Standardization of Neural Networks Verification with VNNLIB and CoCoNet. pages 47–34.

[9] Dorin Doncenco, Julien Girard-Satabin, Romain Xu-Darme, and Zakaria Chihani. A dive into formal explainable attributions for image classification. In *European Conference on Artificial Inteligence 2025*, 2025.

[10] Guy Katz, Clark Barrett, David Dill, Kyle Julian, and Mykel Kochenderfer. Reluplex: An Efficient SMT Solver for Verifying Deep Neural Networks.

[11] Haitham Khedr and Yasser Shoukry. CertiFair: A Framework for Certified Global Fairness of Neural Networks.

[12] Augustin Lemesle, Julien Lehmann, Tristan Le Gall, and Zakaria Chihani. Verifying neural networks with pyrat. In Hakjoo Oh and Yulei Sui, editors, *Static Analysis*, pages 11–33, Cham, 2026. Springer Nature Switzerland.

[13] Guido Manfredi and Yannick Jestin. An introduction to ACAS Xu and the challenges ahead. In *2016 IEEE/AIAA 35th Digital Avionics Systems Conference (DASC)*, pages 1–9. IEEE.

[14] Antoine Miné. Tutorial on Static Inference of Numeric Invariants by Abstract Interpretation. 4(3–4):120–372.

[15] Sanjit A. Seshia, Ankush Desai, Tommaso Dreossi, Daniel J. Fremont, Shromona Ghosh, Edward Kim, Sumukh Shivakumar, Marcell Vazquez-Chanlatte, and Xiangyu Yue. Formal Specification for Deep Neural Networks. In Shuvendu K. Lahiri and Chao Wang, editors, *Automated Technology for Verification and Analysis*, volume 11138, pages 20–34. Springer International Publishing.

[16] Gagandeep Singh, Timon Gehr, Matthew Mirman, Markus Püschel, and Martin Vechev. Fast and effective robustness certification. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

[17] Gagandeep Singh, Timon Gehr, Markus Püschel, and Martin Vechev. An abstract domain for certifying neural networks. 3:1–30.

[18] Jules Soria, Zakaria Chihani, Julien Girard-Satabin, Alban Grastien, Romain Xu-Darme, and Daniela Cancila. Formal abductive latent explanations for prototype-based networks. In *The Fortieth AAAI Conference on Artificial Intelligence*, 2026.

[19] Shiqi Wang, Huan Zhang, Kaidi Xu, Xue Lin, Suman Jana, Cho-Jui Hsieh, and J. Zico Kolter. Beta-CROWN: Efficient Bound Propagation with Per-neuron Split Constraints for Complete and Incomplete Neural Network Robustness Verification.