Automatic Abstraction Refinement for the Verification of Neural Network Control Systems

Technical University of Munich

Department of Informatics

Chair of Robotics, Artificial Intelligence and Real-time Systems

Background

Neural networks have gained tremendous importance in various applications, including safety-critical tasks [6]. They are, however, prone to adversarial attacks [2], i.e. small perturbations in the input can lead to very different outputs, which limits the applicability in safety-critical tasks.

A solution is to verify the neural network before executing the proposed action, considering uncertainties such as sensor noise: This is achieved by modeling the state uncertainty as a set and conservatively propagating the set through the neural network. This conservative output set is then further used to verify the proposed action in the controlled system using reachability analysis. Given the considered uncertainties, the proposed action is safe if the computed reachable set fulfills a given specification.

Since sets cannot be evaluated exactly on nonlinear activation functions (e.g., sigmoid), an over-approximation of the output set of the neural network has to be computed. Previous works [3, 4] approximate the nonlinear layers of the neural network using polynomials and bound the approximation error. Higher-order polynomials are more accurate in approximating the nonlinear function, leading to tighter output sets. However, one quickly runs into performance issues. Consequently, it is desirable to use the coarsest abstraction level possible and only refine it if necessary.

Unfortunately, there does not yet exist an approach to apply this automatic refinement approach in neural network control systems, where the network is repeatedly evaluated, and different refinement schemes are necessary for different time steps.

Description

The main focus of this work is the implementation of an automatic abstraction refinement approach for neural network control systems, especially a simulation-based refinement approach should be implemented and evaluated in this work:

We initially only use linear polynomials for approximating the nonlinear layers of a neural network for all time steps. Whenever the network is evaluated, we use random simulations sampled from the reachable set at the current time step as heuristics if a given specification can still be fulfilled. If the simulations show a violation of the specification, a refinement strategy has to be applied on a prior time step. The results from [4] should be integrated. Different approaches for deciding when to start refining the network should be tested and compared.

Tasks

- Literature research of neural network verification
- Familiarize with the toolbox CORA [1]
- Implementation of automatic abstraction refinement approaches for neural network control system
- Evaluate the approaches on benchmarks from the ARCH competition [5]
- Optional: Integrate adaptive tuning for the parameters of the reachability analysis [7]

References

[1] Matthias Althoff. An introduction to cora 2015. In *ARCH@ CPSWeek*, pages 120–151, 2015.

[2] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

Supervisor:
Prof. Dr.-Ing. Matthias Althoff

Advisor:
Tobias Ladner, M.Sc.

Research project:
FAI

Type:
MA

Research area:
Formal verification, neural networks

Programming language:
MATLAB

Required skills:
Knowledge in formal methods and machine learning, good mathematical background

Language:
English

Date of submission:
17. Oktober 2023

For more information please contact us:

Phone: +49 (89) 289 - 18140

E-Mail: tobias.ladner@tum.de

Website: www.ce.cit.tum.de/air/

[3] Niklas Kochdumper, Christian Schilling, Matthias Althoff, and Stanley Bak. Open-and closed-loop neural network verification using polynomial zonotopes. In *NASA Formal Methods Symposium*, pages 16–36. Springer, 2023.

[4] Tobias Ladner and Matthias Althoff. Automatic abstraction refinement in neural network verification using sensitivity analysis. In *Proceedings of the 26th ACM International Conference on Hybrid Systems: Computation and Control*, pages 1–13, 2023.

[5] Diego Manzanas Lopez, Matthias Althoff, Luis Benet, Xin Chen, Jiameng Fan, Marcelo Forets, Chao Huang, Taylor T Johnson, Tobias Ladner, Wenchao Li, et al. Arch-comp22 category report: artificial intelligence and neural network control systems (ainncs) for continuous and hybrid systems plants. In *Proceedings of 9th International Workshop on Applied Verification of Continuous and Hybrid Systems (ARCH22)*, 2022.

[6] Debasmita Mukherjee, Kashish Gupta, Li Hsin Chang, and Homayoun Najjaran. A survey of robot learning strategies for human-robot collaboration in industrial settings. *Robotics and Computer-Integrated Manufacturing*, 73, 2022.

[7] Mark Wetzlinger, Niklas Kochdumper, and Matthias Althoff. Adaptive parameter tuning for reachability analysis of linear systems. In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 5145–5152. IEEE, 2020.

Technical University of Munich

Department of Informatics

Chair of Robotics, Artificial Intelligence and Real-time Systems