Tutorial: Safe, Secure, and Trustworthy Artificial Intelligence (AI) via Formal Verification of Neural Networks and Autonomous Cyber-Physical Systems (CPS) with NNV

Taylor T. Johnson
Computer Science
Vanderbilt University
Nashville, TN
taylor.johnson@vanderbilt.edu

Diego Manzanas Lopez
Institute for Software Integrated Systems
Vanderbilt University
Nashville, TN
diego.manzanas.lopez@vanderbilt.edu

Hoang-Dung Tran
Computer Science and Engineering
University of Nebraska Lincoln
Lincoln, NE
trhoangdung@gmail.com

***Abstract*—Ensuring safe, secure, and trustworthy artificial intelligence (AI), particularly within safety-critical systems like autonomous cyber-physical stems (CPS), is of paramount importance and of crucial urgency for dependability research. One approach to establishing such desiderata of AI is through formal verification, particularly in machine learning (ML) components like neural networks, to establish they meet certain formal specifications. The Neural Network Verification (NNV) software tool implements automated formal methods for this purpose, specifically reachability analysis, and this interactive tutorial will demonstrate these to formally verify specifications in neural networks, as well as in closed-loop CPS. The tutorial begins with a lecture on the emerging research area of neural network verification, followed by interactive demos of these methods implemented in NNV. Examples will be shown from the security, medicine, and CPS domains.**

***Index Terms*—trustworthy AI, formal methods, cyber-physical systems**

I. Overview

As artificial intelligence (AI) components and especially data-driven machine learning (ML) components are increasingly deployed across domains, establishing desiderata of such components such as their adherence to safety and security specifications is essential to enable trustworthy AI. It is well known certain ML components, particularly deep neural networks (DNNs), suffer from issues such as lack of robustness and susceptibility to adversarial perturbations, where small changes in inputs may lead to drastically different outputs, and further, such systems often just do not function as intended [1]. As one approach to show DNNs meet specifications, neural network verification aims to prove properties of neural networks, in line with formal methods and automated formal verification for other classes of systems, such as digital logic circuits with classical model checking, albeit other challenges arise with AI and ML components. NNV [2], [3] is a software tool[1] for verifying neural networks of various types, along with their usage in autonomous cyber-physical systems (CPS), specifically for a class of closed-loop systems known as neural network control systems (NNCS), where a neural network is used as a feedback controller. There is an emerging research community in neural network [4] and NNCS verification [5], with growing numbers of publications, software tools, workshops/symposia, and competitions like the Applied Verification for Continuous and Hybrid Systems Competition (ARCH-COMP) category on AI and NNCS (AINNCS) and the Verification of Neural Networks Competition (VNN-COMP), as well as with emerging standard formats for neural networks like ONNX[2] and specification languages like VNN-LIB[3]. Through this tutorial, we provide an introduction to these research areas and illustrate on interactive demonstrations and examples the potential of formal verification to establish the safe, secure, and trustworthy development of AI.

[1]https://github.com/verivital/nnv

II. Tutorial Plan

The tutorial is divided into three main sections. It begins with a lecture overview of what is safe, secure, and trustworthy AI, how formal verification can be used to establish specifications, and an introduction to formal verification of neural networks and NNCS. Following this introductory lecture, we move to two hands-on tutorials using NNV for (1) neural network verification and (2) autonomous CPS verification[4].

Next, we provide a more detailed overview of the three planned sections.

a) Overview: To begin the tutorial, we motivate why safe, secure, and trustworthy AI are important, particularly in the context of safety-critical systems, such as autonomous CPS, that increasingly rely on ML components. We next define the neural network verification problem, followed by an overview of various approaches for it through methods developed within the research community, such as optimization, reachability, and SMT-based approaches [4]. In essence,

[2]https://onnx.ai/
[3]https://www.vnnlib.org/
[4]https://github.com/verivital/nnv/tree/master/code/nnv/examples/Tutorial

2833-292X/24/$31.00 ©2024 IEEE
DOI 10.1109/DSN-S60304.2024.00027

the neural network verification problem considers a neural network represented as a function $f : \mathbb{R}^n \mapsto \mathbb{R}^m$, then for a subset of the input space $X \subseteq \mathbb{R}^n$ computes (exactly or overapproximately) the image (range) of f under X, so the set of outputs is $Y = f(X)$. Given a set of states in the output space $B \subseteq \mathbb{R}^m$ representing an undesired behavior, one then checks whether $B \cap f(X) = \emptyset$ or not to determine whether there exists a point $x \in X$ that the neural network can take into B. While underlying approaches vary, this formulation is in essence what most neural network verification methods perform, with many details to be handled regarding what the underlying layer types do to transform inputs.

b) Neural Network Verification: Building on these fundamentals, the next part of the tutorial demonstrates NNV on several tasks, including computer vision problems emerging from the CPS and medical domains (classification, image recognition, and semantic segmentation) and security problems (malware classification) [6]–[10]. Through these demonstrations, we show how to evaluate the robustness of neural networks against targeted and random adversarial attacks, as well as to prove broader specifications. The tutorial includes full step-by-step instructions for how to create and load models, as well as define and analyze specifications with NNV, including an overview of different modes of operation and parameters, along with design guidance and recommendations, such as what layer types are supported and which are complex to analyze (e.g., minimize the number of rectified linear unit [ReLU] layers and total number of ReLU neurons).

c) Autonomous CPS Verification: For the final part of the tutorial, we show how to verify autonomous CPS where a neural network is used as a feedback controller for a plant model, with examples from domains such as autonomous vehicles in different operating domains [5], [11]–[14]. This part will discuss details such as plant modeling (as ordinary differential equations [ODEs] or generalization thereof like hybrid automata), and demonstrate reachability analysis results for such systems, along with a discussion of broader model classes like neural ODEs [15].

d) Summary: The tutorial concludes with a summary of what was presented and demonstrated with NNV and exciting directions for future research to further enable the vision of safe, secure, and trustworthy AI.

Acknowledgments

The material presented in this tutorial is based upon work supported by the National Science Foundation (NSF) through grant numbers 2028001, 2220418, 2220426, 2220401, and 2331937, and the NSF Nebraska EPSCoR under grant OIA-2044049, the Defense Advanced Research Projects Agency (DARPA) under contract numbers FA8750-18-C-0089 and FA8750-23-C-0518, and the Air Force Office of Scientific Research (AFOSR) under contract numbers FA9550-22-1-0019 and FA9550-23-1-0135. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of AFOSR, DARPA, or NSF.

REFERENCES

[1] I. D. Raji, I. E. Kumar, A. Horowitz, and A. Selbst, "The fallacy of ai functionality," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '22. New York, NY, USA: Association for Computing Machinery, 2022, p. 959–972.

[2] H.-D. Tran, X. Yang, D. Manzanas Lopez, P. Musau, L. V. Nguyen, W. Xiang, S. Bak, and T. T. Johnson, "Nnv: The neural network verification tool for deep neural networks and learning-enabled cyber-physical systems," in *Computer Aided Verification*, S. K. Lahiri and C. Wang, Eds. Cham: Springer International Publishing, 2020, pp. 3–17.

[3] D. M. Lopez, S. W. Choi, H.-D. Tran, and T. T. Johnson, "Nnv 2.0: The neural network verification tool," in *Computer Aided Verification*, C. Enea and A. Lal, Eds. Cham: Springer Nature Switzerland, 2023, pp. 397–412.

[4] C. Brix, M. N. Müller, S. Bak, T. T. Johnson, and C. Liu, "First three years of the international verification of neural networks competition (vnn-comp)," *International Journal on Software Tools for Technology Transfer*, vol. 25, no. 3, pp. 329–339, Jun 2023.

[5] D. M. Lopez, M. Althoff, L. Benet, X. Chen, J. Fan, M. Forets, C. Huang, T. T. Johnson, T. Ladner, W. Li, C. Schilling, and Q. Zhu, "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)*, ser. EPiC Series in Computing, G. Frehse, M. Althoff, E. Schoitsch, and J. Guiochet, Eds., vol. 90. EasyChair, 2022, pp. 142–184.

[6] H.-D. Tran, S. Bak, W. Xiang, and T. T. Johnson, "Verification of deep convolutional neural networks using imagestars," in *Computer Aided Verification*, S. K. Lahiri and C. Wang, Eds. Cham: Springer International Publishing, 2020, pp. 18–42.

[7] H.-D. Tran, N. Pal, P. Musau, D. M. Lopez, N. Hamilton, X. Yang, S. Bak, and T. T. Johnson, "Robustness verification of semantic segmentation neural networks using relaxed reachability," in *Computer Aided Verification*, A. Silva and K. R. M. Leino, Eds. Cham: Springer International Publishing, 2021, pp. 263–286.

[8] P. K. Robinette, D. M. Lopez, and T. T. Johnson, "Benchmark: Neural network malware classification," in *Bridging the Gap Between AI and Reality*, B. Steffen, Ed. Cham: Springer Nature Switzerland, 2024, pp. 291–298.

[9] D. Manzanas Lopez and T. T. Johnson, "Empirical analysis of benchmark generation for the verification of neural network image classifiers," in *Bridging the Gap Between AI and Reality*, B. Steffen, Ed. Cham: Springer Nature Switzerland, 2024, pp. 331–347.

[10] P. K. Robinette, D. M. Lopez, S. Serbinowska, K. Leach, and T. T. Johnson, "Case study: Neural network malware detection verification for feature and image datasets," in *International Conference on Formal Methods in Software Engineering (FormaliSE)*, 2024.

[11] H.-D. Tran, F. Cai, M. L. Diego, P. Musau, T. T. Johnson, and X. Koutsoukos, "Safety verification of cyber-physical systems with reinforcement learning control," *ACM Trans. Embed. Comput. Syst.*, vol. 18, no. 5s, oct 2019.

[12] W. Xiang, H. D. Tran, X. Yang, and T. T. Johnson, "Reachable set estimation for neural network control systems: A simulation-guided approach," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–10, 2020.

[13] D. M. Lopez, M. Althoff, M. Forets, T. T. Johnson, T. Ladner, and C. Schilling, "Arch-comp23 category report: Artificial intelligence and neural network control systems (ainncs) for continuous and hybrid systems plants," in *Proceedings of 10th International Workshop on Applied Verification of Continuous and Hybrid Systems (ARCH23)*, ser. EPiC Series in Computing, G. Frehse and M. Althoff, Eds., vol. 96. EasyChair, 2023, pp. 89–125.

[14] D. Manzanas Lopez, T. T. Johnson, S. Bak, H.-D. Tran, and K. L. Hobbs, "Evaluation of neural network verification methods for air-to-air collision avoidance," *Journal of Air Transportation*, vol. 31, no. 1, pp. 1–17, 2023.

[15] D. Manzanas Lopez, P. Musau, N. Hamilton, and T. Johnson, "Reachability analysis of a general class of neural ordinary differential equation," in *Proceedings of the 20th International Conference on Formal Modeling and Analysis of Timed Systems (FORMATS 2022), Co-Located with CONCUR, FMICS, and QEST as part of CONFEST 2022.*, Warsaw, Poland, September 2022.