Tutorial: Neural Network and Autonomous Cyber-Physical Systems Formal Verification for Trustworthy AI and Safe Autonomy

Hoang-Dung Tran
trhoangdung@gmail.com
University of Nebraska-Lincoln
Lincoln, NE, USA

Diego Manzanas Lopez
diego.manzanas.lopez@vanderbilt.edu
Vanderbilt University
Nashville, TN, USA

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

ABSTRACT

This interactive tutorial describes state-of-the-art methods for formally verifying neural networks and their usage within safety-critical cyber-physical systems (CPS). The inclusion of deep learning models in safety-critical applications requires to formally analyze the behavior of the system, including reasoning about the individual components (e.g., controller robustness), and their interactions and effects in the system as a whole. This tutorial begins with a lecture on this emerging research area, followed by demos of these methods implemented in software tools, specifically the Neural Network Verification (NNV) tool. Examples include systems from aerospace, automotive, and beyond.

1 INTRODUCTION

NNV (*N*eural *N*etwork *V*erification) is a software tool[1] that supports the verification of multiple deep learning models, as well as learning-enabled CPS, specifically a class of systems known as Neural Network Control Systems (NNCS) [6, 19], where a neural network is used as a feedback controller in a closed-loop system. The center of NNV is reachability algorithms and various set representations such as star sets, polytopes, zonotopes, and ImageStars, which provide the ability to compute exact and over-approximate reachable sets of feedforward neural networks (FFNN) [15, 16, 19], Convolutional Neural Networks (CNN) [17], Recurrent Neural Networks (RNNs) [14], Semantic Segmentation Neural Networks SSNNs (encoder-decoder architectures) [18], Binary Neural Networks (BNNs) [4], Neural Ordinary Differential Equations [10] and NNCS [8, 13, 19]. The constructed reachable sets can be used to verify various specifications such as safety or robustness, which are commonly used in learning-enabled CPS and deep learning domains, respectively [5, 11].

In the past few years, there has been increased interest in the areas of neural network [1, 11] and NNCS verification [5], not only growing in the number of publications, competitions like VNN-COMP and ARCH-COMP AINNCS, software tools and improved verification methods, but also maturing as a field with standard formats for neural networks like ONNX [12] and specifications in VNN-LIB [3]. This growing interest has also reached outside of academia, leading to the creation of companies as spin-offs from their research [2]. The growing interest has also increased the usage of NNV in several institutions outside of the developers' organizations, including AFRL, Collins Aerospace [2], Northrop Grumman, General Motors, and Toyota.

In this interactive tutorial, we demonstrate NNV capabilities through a collection of safety and robustness verification tasks, which involve the reachable set computation of feedforward, convolutional, semantic segmentation, and recurrent neural networks, as well as neural ordinary differential equations and neural network control systems. And as NNV is publicly available, participants can follow along as desired. Publications on NNV have participated in several prior repeatability/artifact evaluations at top conferences such as CAV [6, 19], with passing results for multiple publications, which illustrates the feasibility of this interactive demonstration plan. Further, NNV is already available for in-browser execution through platforms like CodeOcean [7], and we plan to organize the interactive tutorial aspects around this or similar (Jupyter-like) in-browser demonstrations.

Based on the growing importance of safety, trust, and trustworthiness in AI and especially in safety-critical autonomous CPS, we imagine this tutorial will be of interest to the attendees. In particular, we believe the tutorial is timely and relevant for ESWeek and EMSOFT given the focus of our work on developing NNV in the context of the typical embedded and cyber-physical model-based design flow using tools like Matlab and Simulink.

2 PRESENTATION FORMAT AND TUTORIAL PLAN

The tutorial is divided into three main sections, beginning with an overview of formal verification, safe autonomy and trustworthy AI, and an introduction to formal verification of neural networks, followed by two hands-on tutorials using NNV for neural network verification and autonomous CPS verification. The anticipated timeframe for the tutorial is a half day, with approximately an hour devoted to each of these three sections. The planned presenters are included next in the tentative agenda. If time allows, we will include a discussion period at the end for around fifteen minutes, and we will take questions throughout the tutorial. We will accommodate any planned coffee breaks, etc. based on the program agenda as it is finalized.

(1) Overview (motivation, safe autonomy, trustworthy AI, formal verification of neural networks and NNCS): *Taylor T. Johnson*

[1]https://github.com/verivital/nnv
[2]https://datenvorsprung.at/, https://latticeflow.ai/

(2) Neural network verification (open loop tasks for CNNs, SSNNs, BNNs, etc.): *Hoang-Dung Tran*
(3) Autonomous CPS verification (closed loop / NNCS): *Diego Manzanas Lopez*

Next, we include some further detail on the planned topics in these three portions.

Overview. In this portion of the tutorial, we first motivate why safe autonomy and trustworthy AI are important, particularly in the context of autonomous CPS that incorporate machine learning components. We then discuss what neural network verification is, surveying the various approaches for it using different methods developed within the emerging field (such as optimization and SMT-based approaches, beyond our reachability approaches in NNV), and preview important and impactful use cases in the embedded systems industry and high-profile research programs, such as DARPA Assured Autonomy and ANSR, as well as NSF Safe Learning-Enabled Systems, as well as within the research community through VNN-COMP and ARCH-COMP AINNCS.

Neural Network Verification. In this portion of the tutorial, we demonstrate the capabilities of NNV in a wide variety of open-loop applications including classification, image recognition, and semantic segmentation over a collection of deep learning models such as CNNs [17], RNNs [14], and SSNNs [18]. Throughout these experiments, we evaluate the robustness of the neural network models against targeted and random adversarial attacks to the system.

Autonomous CPS verification. In this portion of the tutorial, we focus on the verification of autonomous CPS in safety-critical applications, including examples in aerospace, ground, and maritime autonomous vehicles [5, 8, 9, 13]. We present an interactive tutorial that shows step by step how lo load and create a NNCS model in NNV, to represent the specifications to verify, compute the reachable sets of the system, and finally show the verification proofs or counterexamples as well as the visualization of the computed reachable sets.

ACKNOWLEDGMENTS

The material presented in this paper is based upon work supported by the National Science Foundation (NSF) through grant numbers 1910017, 2028001, 2220418, 2220426 and 2220401, 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] Christopher Brix, Mark Niklas Müller, Stanley Bak, Taylor T. Johnson, and Changliu Liu. 2023. First Three Years of the International Verification of Neural Networks Competition (VNN-COMP). arXiv:2301.05815 [cs.LG]
[2] Arthur Clavière, Laura Altieri Sambartolomé, Eric Asselin, Christophe Garion, and Claire Pagetti. 2022. Verification of Machine Learning Based Cyber-Physical Systems: A Comparative Study. In *25th ACM International Conference on Hybrid Systems: Computation and Control* (Milan, Italy) *(HSCC '22)*. Association for Computing Machinery, New York, NY, USA, Article 22, 16 pages. https://doi.org/10.1145/3501710.3519540
[3] Dario Guidotti, Stefano Demarchi, Armando Tacchella, and Luca Pulina. [n. d.]. *The Verification of Neural Networks Library (VNN-LIB).* https://www.vnnlib.org, 2022
[4] Michael Ivashchenko, Sungwoo Choi, Viet-Luan Nguyen, and Hoang-Dung Tran. 2023. Verifying Binary Neural Networks on Continuous Input Space using Star Reachability. In *International Conference on Formal Methods in Software Engineering*. ACM.
[5] Diego Manzanas Lopez, Matthias Althoff, Luis Benet, Xin Chen, Jiameng Fan, Marcelo Forets, Chao Huang, Taylor T Johnson, Tobias Ladner, Wenchao Li, Christian Schilling, and Qi Zhu. 2022. 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) (EPiC Series in Computing, Vol. 90)*, Goran Frehse, Matthias Althoff, Erwin Schoitsch, and Jeremie Guiochet (Eds.). EasyChair, 142–184.
[6] Diego Manzanas Lopez, Sung Woo Choi, Hoang-Dung Tran, and Taylor T. Johnson. 2023. NNV 2.0: The Neural Network Verification Tool. In *35th International Conference on Computer-Aided Verification (CAV).*
[7] Diego Manzanas Lopez, Sung Woo Choi, Hoang-Dung Tran, and Taylor T. Johnson. 2023. NNV 2.0: The Neural Network Verification Tool, (CodeOcean Capsule). https://doi.org/10.24433/CO.0803700.v1. https://doi.org/10.24433/CO.0803700.v1
[8] Diego Manzanas Lopez, Taylor T. Johnson, Stanley Bak, Hoang-Dung Tran, and Kerianne Hobbs. 2022. Evaluation of Neural Network Verification Methods for Air to Air Collision Avoidance. *AIAA Journal of Air Transportation (JAT)* (Oct. 2022).
[9] Diego Manzanas Lopez, Patrick Musau, Nathaniel Hamilton, Hoang-Dung Tran, and Taylor T. Johnson. 2020. Case Study: Safety Verification of an Unmanned Underwater Vehicle. In *Workshop on Assured Autonomous Systems (WAAS).* IEEE.
[10] Diego Manzanas Lopez, Patrick Musau, Nathaniel Hamilton, and Taylor Johnson. 2022. 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.
[11] Mark Niklas Müller, Christopher Brix, Stanley Bak, Changliu Liu, and Taylor T. Johnson. 2022. The Third International Verification of Neural Networks Competition (VNN-COMP 2022): Summary and Results.
[12] ONNX. [n. d.]. *Open Neural Network Exchange (ONNX).* https://onnx.ai/
[13] Hoang-Dung Tran, Feiyang Cei, Diego Manzanas Lopez, Taylor T. Johnson, and Xenofon Koutsoukos. 2019. Safety Verification of Cyber-Physical Systems with Reinforcement Learning Control. In *ACM SIGBED International Conference on Embedded Software (EMSOFT'19).* ACM.
[14] Hoang Dung Tran, SungWoo Choi, Tomoya Yamaguchi, Bardh Hoxha, and Danil Prokhorov. 2023. Verification of Recurrent Neural Networks using Star Reachability. In *The 26th ACM International Conference on Hybrid Systems: Computation and Control (HSCC).*
[15] Hoang-Dung Tran, Patrick Musau, Diego Manzanas Lopez, Xiaodong Yang, Luan Viet Nguyen, Weiming Xiang, and Taylor T. Johnson. 2019. Parallelizable Reachability Analysis Algorithms for Feed-forward Neural Networks. In *Proceedings of the 7th International Workshop on Formal Methods in Software Engineering (FormaliSE'19)* (Montreal, Quebec, Canada) *(FormaliSE '19).* IEEE Press, Piscataway, NJ, USA, 31–40. https://doi.org/10.1109/FormaliSE.2019.00012
[16] Hoang-Dung Tran, Patrick Musau, Diego Manzanas Lopez, Xiaodong Yang, Luan Viet Nguyen, Weiming Xiang, and Taylor T. Johnson. 2019. Star-Based Reachability Analsysis for Deep Neural Networks. In *23rd International Symposium on Formal Methods (FM'19).* Springer International Publishing.
[17] Hoang-Dung Tran, Neelanjana Pal, Diego Manzanas Lopez, Patrick Musau, Xiaodong Yang, Luan Viet Nguyen, Weiming Xiang, Stanley Bak, and Taylor T Johnson. 2021. Verification of piecewise deep neural networks: a star set approach with zonotope pre-filter. *Formal Aspects of Computing* 33, 4 (2021), 519–545.
[18] Hoang-Dung Tran, Neelanjana Pal, Patrick Musau, Diego Manzanas Lopez, Nathaniel Hamilton, Xiaodong Yang, Stanley Bak, and Taylor T Johnson. 2021. Robustness verification of semantic segmentation neural networks using relaxed reachability. In *International Conference on Computer Aided Verification.* Springer, 263–286.
[19] Hoang-Dung Tran, Xiaodong Yang, Diego Manzanas Lopez, Patrick Musau, Luan Viet Nguyen, Weiming Xiang, Stanley Bak, and Taylor T. Johnson. 2020. NNV: The Neural Network Verification Tool for Deep Neural Networks and Learning-Enabled Cyber-Physical Systems. In *32nd International Conference on Computer-Aided Verification (CAV).*