Track C1: Safety Verification of Deep Neural Networks (DNNs)

Daniel Neider[1,2][0000−0001−9276−6342] and Taylor T. Johnson[3][0000−0001−8021−9923]

[1] Chair for Verification and Formal Guarantees of Machine Learning, TU Dortmund University, Dortmund, Germany
daniel.neider@tu-dortmund.de
[2] Research Center for Trustworthy Data Science and Security, Dortmund, Germany
[3] Institute for Software Integrated Systems, Vanderbilt University, Nashville, TN, USA
taylor.johnson@vanderbilt.edu

Abstract. Formal verification of neural networks and broader machine learning models is an emerging field that has gained significant attention due to the growing use and impact of these data-driven methods. This track explores techniques for formally verifying neural networks and other machine learning models across various application domains. It includes papers and presentations discussing new methodologies, software frameworks, technical approaches, and case studies. Benchmarks play a crucial role in evaluating the effectiveness and scalability of these methods. Currently, available benchmarks mainly focus on computer vision problems, such as local robustness to adversarial perturbations of image classifiers. To address this limitation, this track compiles and publishes benchmarks comprising machine learning models and their specifications across domains such as computer vision, finance, security, and others. These benchmarks will help assess the suitability and applicability of formal verification methods in diverse domains.

Keywords: Formal Verification · Formal Methods · Neural Networks · Safety of Autonomy.

1 Description of the Track

Formal verification of deep neural networks (DNNs) and broader machine learning models has been a burgeoning field in the past few years, with continued increasing interest given the ongoing growth and applicability of these data-driven methods. This track focuses on methods for formal verification of machine learning models, including neural networks, but also beyond to other model types across application domains. In particular, this track features contributions addressing theoretical aspects of verifying neural networks [5], the dev-ops cycle for trustworthy learning-enabled autonomous systems [2], neural networks related

to cyber-physical systems [21,4], neuro-symbolic verification [22], anomaly detection [9], predictive maintenance [12], and the problem of overconfident neural networks [8].

In addition, benchmarks are critical for evaluating scalability and broader progress within formal methods. Most recent benchmarks for evaluating neural network verification methods and broader machine learning verification have focused predominantly on computer vision problems, specifically local robustness to adversarial perturbations of image classifiers. However, neural networks and machine learning models are being used across a variety of safety and security-critical domains, and domain-specific benchmarks—both in terms of the machine learning models and their specifications—are necessary to identify limitations and directions for improvements, as well as to evaluate and ensure applicability of these methods in these domains. For instance, controllers in autonomous systems are increasingly created with data-driven methods, and malware classifiers in security are often neural networks, each of which domain has its specificities, as do countless other applications in cyber-physical systems, finance, science, and beyond.

Our second contribution involves compiling and publishing benchmarks comprising models and specifications across various domains where formal verification of neural networks and machine learning is being explored. These benchmarks span several fields, including information technology (IT) security [20], computer vision [19,14], autonomous driving [16], aerospace [10], maritime search and rescue [11], and chemical process engineering [17]. They have been designed to serve as inputs for future iterations of the International Verification of Neural Networks Competition (VNN-COMP) [1,18,3] and the Artificial Intelligence and Neural Network Control Systems (AINNCS) category [15,7,6,13] of the International Competition on Verifying Continuous and Hybrid Systems (ARCH-COMP). By sharing these benchmarks with the scientific community, our goal is to encourage further research and inspire conversations about expanding the scope of the Verification of Neural Networks standard (VNN-LIB).

As indicated above, the contributions of this track can be grouped into two categories: (i) verification of neural networks and autonomous systems and (ii) verification benchmarks. The remainder of this article gives an overview of each contribution in this track. In the spirit of AISoLA's aims and scope, we have utilized ChatGPT to help summarize the contributions, specifically utilizing the August 3rd, 2023 release of GPT-4 and code interpreter to extract the paper contents from the provided PDFs, to generate a few sentence overview of each contribution, which we have manually reviewed and edited for accuracy. We have organized the contributions according to the planned sessions during the event, in essence into verification approaches and benchmarks.

2 Verification of Neural Networks and Autonomous Systems

Papers and presentations on verification approaches for neural networks are in contributions [12,5,22,8,9] and for autonomous systems in [21,2,4], each of which is briefly summarized next.

The paper, "Formal Verification of a Neural Network Based Prognostics System for Aircraft Equipment," [12] presents methodology for verifying properties of a deep convolutional neural network (CNN) used for estimating the remaining useful life (RUL) of aircraft mechanical equipment. The authors provide mathematical formalizations of the estimator requirements, such as stability and monotonicity, and encode these properties as linear constraints. They use a state-of-the-art tool for neural network verification to check these properties on a neural network model of a prognostics system trained on a real-world dataset of bearing degradation data.

The paper, "The inverse problem for neural networks," [5] investigates the problem of computing the preimage of a set under a neural network with piecewise-affine activation functions. The authors revisit an old result that the preimage of a polyhedral set is again a union of polyhedral sets and can be effectively computed. They show several applications of computing the preimage for analysis and interpretability of neural networks. This study is essential for understanding the inverse problem for neural networks and its implications on the analysis and interpretability of neural networks.

The paper, "Distribution-Aware Neuro-Symbolic Verification," [22] proposes a novel approach to restrict verification processes to the data distribution (or other distributions) by using distribution-aware neuro-symbolic verification. The authors propose the non-linear independent component estimator (NICE) Laplacianizing flow as a suitable density model due to its two major properties: the associated log-density function is piece-wise affine, making it applicable for SMT based approaches using linear arithmetic, and each NICE flow maps the upper log-density level sets of the data distribution to the upper log-density level sets of the latent Laplacian, which has potential applications within interval bound propagation methods.

The contribution, "Towards Formal Guarantees for Networks' Overconfidence" [8] addresses the problem of neural networks making overconfident predictions when presented with out-of-distribution inputs. The authors note that even well-trained networks can show very high confidence for inputs that do not belong to the task they are trained for, raising concerns about real-world scenarios where a wide range of inputs may be encountered. They discuss several existing works that suggest training approaches aimed at decreasing the confidence of out-of-distribution inputs, mainly by regularization terms. Each of these works considers a certain type of out-of-distribution domain, samples from it, and incorporates the samples into the training process.

The paper, "Towards Verification of Changes in Dynamic Machine Learning Models using Deep Ensemble Anomaly Detection," [9] emphasizes the importance of formal verification of machine learning models in safety-critical systems,

especially when dealing with dynamically changing data distributions. The authors propose considering the temporal representation of state changes as the context of a dynamic machine learning model, and illustrate this with environmental time series variables such as temperature, humidity, and illumination used to train models with respect to seasonal states like spring, summer, autumn, and winter. This approach helps in verifying a time-dependent machine learning model with respect to distinct states of an evolving time series.

The paper, "Continuous Engineering for Trustworthy Learning-enabled Autonomous Systems," [2] discusses the challenges and approaches in engineering trustworthy learning-enabled autonomous systems. The paper discusses the importance of continuous engineering in ensuring the trustworthiness of autonomous systems that are enabled by machine learning, and methodology for establishing this. The work involves a collaborative effort from experts in various fields related to autonomous systems, machine learning, and continuous engineering.

Lastly, two presentation-only contributions describing verification methods were given, both targeting autonomous systems. In the talk "Reachability for neural-network control systems" [21], methods for reachability analysis of neural network control systems were presented. In the talk "Verification of a Neural Network for Modelling the Dynamics of a Quadcopter" [4], verification methods for dynamical systems, specifically a quadcopter modeled as a neural network, are presented.

3 Verification Benchmarks

The development of benchmarks for use by the research community is a critical task and aids in a variety of objectives ranging from standardization of formats to identification of critical scalability issues. Through this portion of the track, several benchmarks from a diverse set of problem domains were proposed, specifically in papers [16,20,10,11,17,19,14], each of which is briefly summarized next.

The paper, "Benchmarks: Semantic Segmentation Neural Network Verification and Objection Detection Neural Network Verification in Perceptions Tasks of Autonomous Driving," [16] addresses a critical gap in the field of autonomous driving. Although there are existing benchmarks for verifying the robustness of neural networks, there are hardly any related to autonomous driving, especially those related to object detection and semantic segmentation. The authors present an innovative approach to benchmark formal verification tools and approaches for neural networks in the context of perception tasks of autonomous driving. Specifically, the authors contribute two novel benchmarks: one for semantic segmentation neural network verification and another for object detection neural network verification.

The paper, "Benchmark: Neural Network Malware Classification," [20] addresses the increasing complexity and sophistication of malware threats and the need for advanced detection methods, such as deep neural networks (DNNs), for

malware classification. The authors propose two malware classification benchmarks: a feature-based benchmark and an image-based benchmark. Feature-based datasets provide a detailed understanding of malware characteristics, while image-based datasets transform raw malware binary data into grayscale images for swift processing. These datasets can be used for both binary classification (benign vs. malicious) and classifying known malware into a particular family.

The paper, "Benchmark: Remaining Useful Life Predictor for Aircraft Equipment," [10] proposes a predictive maintenance application as a benchmark problem for the verification of neural networks (VNN). The authors implement a deep learning-based estimator of remaining useful life (RUL) of aircraft mechanical components, such as bearings, as a convolutional neural network. They provide mathematical formalizations of its non-functional requirements, such as stability and monotonicity, as properties. These properties can be used to assess the applicability and scalability of existing VNN tools. The benchmark materials, such as trained models, examples of properties, test datasets, and property generation procedures, are available on GitHub.

The paper, "Benchmark: Object Detection for Maritime Search and Rescue," [11] proposes an object detection system for maritime search and rescue as a benchmark problem for the verification of neural networks (VNN). The model to be verified is a YOLO (You Only Look Once) deep neural network for object detection and classification and has a very high number of learnable parameters (millions). The authors describe the workflow for defining and generating robustness properties in the regions of interest of the images, i.e., in the neighborhood of the objects to be detected by the neural network. This benchmark can be used to assess the applicability and scalability of existing VNN tools for perception systems based on deep learning.

The paper, "Benchmark: Neural Networks for Anomaly Detection in Batch Distillation," [17] presents a benchmark suite for verifying neural networks used in anomaly detection for batch distillation chemical processes. The authors highlight the importance of these models working safely and reliably in safety-critical applications, such as chemical plants, where failure to report anomalies or false alarms may result in hazards to the environment, harm to human life, or substantial financial or scientific loss. The work aims to contribute to the field by providing a benchmark suite that can help in verifying the safety and reliability of neural networks used in such critical applications.

The paper, "Benchmark: Formal Verification of Semantic Segmentation Neural Networks," [19] discusses the application of formal verification techniques to semantic segmentation neural networks. While significant progress has been made in verification methods for various deep neural networks (DNNs), such as feed-forward neural networks (FFNNs) and convolutional neural networks (CNNs), the application of these techniques to semantic segmentation remains largely unexplored. Semantic segmentation networks are crucial in computer vision applications, where they assign semantic labels to individual pixels within an image. Given their deployment in safety-critical domains, ensuring the correctness of these networks becomes paramount.

The paper, "Empirical Analysis of Benchmark Generation for the Verification of Neural Network Image Classifiers," [14] addresses the increasing usage of deep learning technology in safety-critical applications such as autonomous cars and medicine. The authors emphasize that the use of models, e.g., neural networks, in safety-critical applications demands a thorough evaluation from both a component and system-level perspective. Despite great efforts in the formal verification of neural networks in the past decade, several challenges remain, one of which is the development of neural networks for easier verification. The authors aim to address this challenge and contribute to the ongoing efforts in the formal verification of neural networks.

4 Summary and Outlook

This track presents approaches for verification of neural networks and autonomous systems, as well as collecting benchmarks for these types of approaches to be used in the broader research community. In the future, these benchmarks and approaches may be utilized in verification tools participating in events such as VNN-COMP and the ARCH-COMP AINNCS category, and we hope this initiative toward collecting benchmarks continues.

Acknowledgements

The material presented in this paper is based upon work supported by the National Science Foundation (NSF) through grant numbers 2220426 and 2220401, and the Defense Advanced Research Projects Agency (DARPA) under contract number 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, conclusions, or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of AFOSR, DARPA, or NSF. In addition, the work was supported by the Deutsche Forschungsgemeinschaft (DFG) through grant number 459419731.

References

1. Bak, S., Liu, C., Johnson, T.: The second international verification of neural networks competition (VNN-COMP 2021): Summary and results (2021). https://doi.org/10.48550/ARXIV.2109.00498, https://arxiv.org/abs/2109.00498
2. Bensalem, S., Katsaros, P., Nickovic, D., Liao, B.H.C., Nolasco, R.R., Ahmed, M.A.E.S., Beyene, T.A., Cano, F., Delacourt, A., Esen, H., Forrai, A., He, W., Huang, X., Kekatos, N., Konighofer, B., Paulitsch, M., Peled, D., Ponchant, M., Sorokin, L., Tong, S., Wu, C.: Continuous engineering for trustworthy learning-enabled autonomous systems. In: AISoLA: International Symposium on Leveraging Applications of Formal Methods. LNCS, vol. In this volume. Springer International Publishing (Oct 2023)

3. Brix, C., Müller, M.N., Bak, S., Johnson, T.T., Liu, C.: First three years of the international verification of neural networks competition (vnn-comp). International Journal on Software Tools for Technology Transfer pp. 1–11 (2023)
4. Dulai, A., Garcia, L.: Presentation: Verification of a neural network for modelling the dynamics of a quadcopter. In: AISoLA: International Symposium on Leveraging Applications of Formal Methods. LNCS, vol. In this volume. Springer International Publishing (Oct 2023)
5. Forets, M., Schilling, C.: The inverse problem for neural networks. In: AISoLA: International Symposium on Leveraging Applications of Formal Methods. LNCS, vol. In this volume. Springer International Publishing (Oct 2023)
6. Johnson, T.T., Lopez, D.M., Benet, L., Forets, M., Guadalupe, S., Schilling, C., Ivanov, R., Carpenter, T.J., Weimer, J., Lee, I.: Arch-comp21 category report: Artificial intelligence and neural network control systems (ainncs) for continuous and hybrid systems plants. In: Frehse, G., Althoff, M. (eds.) 8th International Workshop on Applied Verification of Continuous and Hybrid Systems (ARCH21). EPiC Series in Computing, vol. 80, pp. 90–119. EasyChair (2021). https://doi.org/10.29007/kfk9
7. Johnson, T.T., Lopez, D.M., Musau, P., Tran, H.D., Botoeva, E., Leofante, F., Maleki, A., Sidrane, C., Fan, J., Huang, C.: Arch-comp20 category report: Artificial intelligence and neural network control systems (ainncs) for continuous and hybrid systems plants. In: Frehse, G., Althoff, M. (eds.) ARCH20. 7th International Workshop on Applied Verification of Continuous and Hybrid Systems (ARCH20). EPiC Series in Computing, vol. 74, pp. 107–139. EasyChair (2020). https://doi.org/10.29007/9xgv
8. Kabaha, A., Cohen, D.D.: Towards formal guarantees for networks' overconfidence. In: AISoLA: International Symposium on Leveraging Applications of Formal Methods. LNCS, vol. In this volume. Springer International Publishing (Oct 2023)
9. Katzke, T., Li, B., Klüttermann, S., , Müller, E.: Towards verification of changes in dynamic machine learning models using deep ensemble anomaly detection. In: AISoLA: International Symposium on Leveraging Applications of Formal Methods. LNCS, vol. In this volume. Springer International Publishing (Oct 2023)
10. Kirov, D., Rollini, S.F.: Benchmark: Remaining useful life predictor for aircraft equipment. In: AISoLA: International Symposium on Leveraging Applications of Formal Methods. LNCS, vol. In this volume. Springer International Publishing (Oct 2023)
11. Kirov, D., Rollini, S.F., Chandrahas, R., Reddy, S., Chandupatla, Sawant, R.: Benchmark: Object detection for maritime search and rescue. In: AISoLA: International Symposium on Leveraging Applications of Formal Methods. LNCS, vol. In this volume. Springer International Publishing (Oct 2023)
12. Kirov, D., Rollini, S.F., Guglielmo, L.D., , Cofer, D.: Formal verification of a neural network based prognostics system for aircraft equipment. In: AISoLA: International Symposium on Leveraging Applications of Formal Methods. LNCS, vol. In this volume. Springer International Publishing (Oct 2023)
13. Lopez, D.M., Althoff, M., Benet, L., Chen, X., Fan, J., Forets, M., Huang, C., Johnson, T.T., Ladner, T., Li, W., Schilling, C., Zhu, Q.: Arch-comp22 category report: Artificial intelligence and neural network control systems (ainncs) for continuous and hybrid systems plants. In: Frehse, G., Althoff, M., Schoitsch, E., Guiochet, J. (eds.) Proceedings of 9th International Workshop on Applied Verification of Continuous and Hybrid Systems (ARCH22). EPiC Series in Computing, vol. 90, pp. 142–184. EasyChair (2022). https://doi.org/10.29007/wfgr

14. Lopez, D.M., Johnson, T.T.: Empirical analysis of benchmark generation for the verification of neural network image classifiers. In: AISoLA: International Symposium on Leveraging Applications of Formal Methods. LNCS, vol. In this volume. Springer International Publishing (Oct 2023)
15. Lopez, D.M., Musau, P., Tran, H.D., Dutta, S., Carpenter, T.J., Ivanov, R., Johnson, T.T.: Arch-comp19 category report: Artificial intelligence and neural network control systems (ainncs) for continuous and hybrid systems plants. In: Frehse, G., Althoff, M. (eds.) ARCH19. 6th International Workshop on Applied Verification of Continuous and Hybrid Systems. EPiC Series in Computing, vol. 61, pp. 103–119. EasyChair (2019). https://doi.org/10.29007/rgv8
16. Luo, Y., Ma, J., Han, S., Xie, L.: Benchmarks: Semantic segmentation neural network verification and objection detection neural network verification in perceptions tasks of autonomous driving. In: AISoLA: International Symposium on Leveraging Applications of Formal Methods. LNCS, vol. In this volume. Springer International Publishing (Oct 2023)
17. Lutz, S., Neider, D.: Benchmark: Neural networks for anomaly detection in batch distillation. In: AISoLA: International Symposium on Leveraging Applications of Formal Methods. LNCS, vol. In this volume. Springer International Publishing (Oct 2023)
18. Müller, M.N., Brix, C., Bak, S., Liu, C., Johnson, T.T.: The third international verification of neural networks competition (VNN-COMP 2022): Summary and results (2022). https://doi.org/10.48550/arXiv.2212.10376, https://arxiv.org/abs/2212.10376
19. Pal, N., Lee, S., Johnson, T.T.: Benchmark: Formal verification of semantic segmentation neural networks. In: AISoLA: International Symposium on Leveraging Applications of Formal Methods. LNCS, vol. In this volume. Springer International Publishing (Oct 2023)
20. Robinette, P.K., Lopez, D.M., Johnson, T.T.: Benchmark: Neural network malware classification. In: AISoLA: International Symposium on Leveraging Applications of Formal Methods. LNCS, vol. In this volume. Springer International Publishing (Oct 2023)
21. Schilling, C.: Presentation: Reachability for neural-network control systems. In: AISoLA: International Symposium on Leveraging Applications of Formal Methods. LNCS, vol. In this volume. Springer International Publishing (Oct 2023)
22. Zaid, F.A., Diekmann, D., Neider, D.: Distribution-aware neuro-symbolic verification. In: AISoLA: International Symposium on Leveraging Applications of Formal Methods. LNCS, vol. In this volume. Springer International Publishing (Oct 2023)