Prof. Dr. Holger Hoos
Chair for AI Methodology
RWTH Aachen University

Open Project:

Enhancing Neural Network Robustness Evaluation in Medical Diagnostics through Formal Verification

Type:

Bachelor Thesis	☐
Master Thesis	☒
Student Assistant	☐

Daily Supervisors:
Konstantin Kaulen
kaulen@aim.rwth-aachen.de
Theaterstrasse 35-39, 52062 Aachen, Room 310

Matthias König
matthias@tuev-lab.ai
TÜV AI.Lab

Description

In recent years, the remarkable performance of deep neural networks has led to their application across various safety-critical domains; those include aircraft collision avoidance systems [1] as well as the medical domain (see, *e.g.*, [2]) In the medical domain, the reliability of neural networks is crucial, especially when those are employed in systems used for diagnostic purposes.
At the same time, it has been shown that neural networks are susceptible to *adversarial samples*, where slight input modifications result in incorrect outputs. These malicious inputs can be crafted using *adversarial attacks* that typically leverage gradient information of the neural network to find perturbations to an input within a predefined noise model that cause misclassificaton (see, *e.g.*, [3, 4]). However, due to their gradient-based nature, adversarial attacks may not identify every adversarial example. Thus, to accurately assess the robustness of a given neural network, *formal verification techniques* are needed. Those provide mathematical guarantees for the robustness of a given neural network by, *e.g.*, encoding the verification task as a Mixed Integer Programming problem [5] or by bounding the output of the neural network for all inputs permitted by the examined noise model [6]. For an extensive overview of the state of the art in neural network verification see, *e.g.*, [7, 8, 9, 10].
In the medical domain, neural networks may be employed to aid physicians in diagnosing cardiovascular diseases using electrocardiograms (ECG) [11]. In this context, it is crucial that networks are robust against domain-specific changes in the input such as sensor noise stemming from the electrodes used for ECG measuring. One approach to evaluate a model against such input variations is to simulate noise based on mathematical noise models [12]. These noise models are applied to the given data, resulting in perturbed observations on which the model is then tested. However, this approach cannot formally prove that a neural network is resistant against all possible perturbations allowed by the noise model since those are not enumerable [13]. While such guarantees should be of utmost importance in the medical domain, formal verification techniques for neural networks employed in ECG classification remain unexplored to date. However, recently substantial efforts have been made to formally verify problems encountered in realistic application scenarios (see, *e.g*, [14, 15, 16]). Furthermore, there are promising advancements in verifying more complex properties such as geometric transformations (see, *e.g.*, [17, 13, 18]).
Therefore, we propose to use formal verification methods to assess a networks's robustness against specific noise types from the medical domain. Instead of simulating noise, these methods use mathematical proofs

to verify a given property and can thereby provide concrete guarantees regarding the robustness of deep-learning-based ECG classifiers.
In summary, this work aims to investige the potential of formal neural network verification in the medical domain by conducting a case study for deep-learning-based ECG classification. By identifying and incorporating relevant noise models into verifiable formulations, this study seeks to rigorously assess the robustness of state-of-the-art ECG classifiers.

Preliminary Tasks

Students are encouraged to adjust this list to their own ideas:

- Conduct a literature review on the state of the art in deep-learning-based ECG classification, neural network verification, ECG noise models and the modeling of verification properties.
- Formulate a robustness property that describes the robustness of ECG classifiers against domain-specific noise models. This formulation should be verifiable by a state-of-the-art verification tool.
- Conduct experiments to investigate the robustness of ECG classifiers against domain-specific noise models using the formulated robustness property.

References

[1] Kyle D Julian, Mykel J Kochenderfer, and Michael P Owen. "Deep Neural Network Compression for Aircraft Collision Avoidance Systems". In: *Journal of Guidance, Control, and Dynamics* 42.3 (2019), pp. 598–608.

[2] Shenda Hong et al. "HOLMES: Health OnLine Model Ensemble Serving for Deep Learning Models in Intensive Care Units". In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2020, pp. 1614–1624.

[3] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. "Explaining and Harnessing Adversarial Examples". In: *3rd International Conference on Learning Representations, (ICLR 2015)*. 2015, pp. 1–11.

[4] Aleksander Madry et al. "Towards Deep Learning Models Resistant to Adversarial Attacks". In: *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[5] Vincent Tjeng, Kai Y. Xiao, and Russ Tedrake. "Evaluating Robustness of Neural Networks with Mixed Integer Programming". In: *Proceedings of the 7th International Conference on Learning Representations (ICLR 2019)*. 2019, pp. 1–21.

[6] Shiqi Wang et al. "Beta-CROWN: Efficient Bound Propagation with Per-neuron Split Constraints for Neural Network Robustness Verification". In: *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*. 2021, pp. 29909–29921.

[7] Matthias König et al. "Critically assessing the state of the art in neural network verification". In: *Journal of Machine Learning Research* 25.12 (2024), pp. 1–53.

[8] Linyi Li, Tao Xie, and Bo Li. "SoK: Certified Robustness for Deep Neural Networks". In: *2023 IEEE Symposium on Security and Privacy (SP)*. May 2023, pp. 1289–1310.

[9] Marta Kwiatkowska and Xiyue Zhang. "When to Trust AI: Advances and Challenges for Certification of Neural Networks". In: *Proceedings of the 18th Conference on Computer Science and Intelligence Systems, FedCSIS 2023, Warsaw, Poland, September 17-20, 2023*. Vol. 35. Annals of Computer Science and Information Systems. 2023, pp. 25–37.

[10] Mark Huasong Meng et al. "Adversarial Robustness of Deep Neural Networks: A Survey from a Formal Verification Perspective". In: *IEEE Transactions on Dependable and Secure Computing* (2022), pp. 1–18.

[11] Nils Strodthoff et al. "Deep Learning for ECG Analysis: Benchmarks and Insights from PTB-XL". In: *IEEE Journal of Biomedical and Health Informatics* 25.5 (2021), pp. 1519–1528.

[12] Bundesamt für Sicherheit in der Informationstechnik. *Einsatz von Künstlicher Intelligenz in medizinischen Diagnose- und Prognosesystemen*. 2024.

[13] Jeet Mohapatra et al. "Towards Verifying Robustness of Neural Networks Against A Family of Semantic Perturbations". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 2020, pp. 241–249.

[14] Cong Liu, Darren D. Cofer, and Denis Osipychev. "Verifying an Aircraft Collision Avoidance Neural Network with Marabou". In: *Proceedings of the 15th NASA Formal Methods Symposium (NFM 2023)*. Vol. 13903. Lecture Notes in Computer Science. Springer, 2023, pp. 79–85.

[15] Patrick Henriksen et al. "Bias Field Robustness Verification of Large Neural Image Classifiers". In: *32nd British Machine Vision Conference 2021, (BMVC 2021)*. BMVA Press, 2021, p. 202.

[16] Andreas Venzke and Spyros Chatzivasileiadis. "Verification of Neural Network Behaviour: Formal Guarantees for Power System Applications". In: *IEEE Transactions on Smart Grid* 12.1 (Jan. 2021), pp. 383–397.

[17] Ben Batten et al. "Verification of Geometric Robustness of Neural Networks via Piecewise Linear Approximation and Lipschitz Optimisation". en. In: *arXiv.org* (Aug. 2024).

[18] Mislav Balunovic et al. "Certifying Geometric Robustness of Neural Networks". In: *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*. 2019, pp. 15287–15297.