Towards Provable Safety in Railway Video Monitoring

Technical University of Munich

Department of Informatics

Chair of Robotics, Artificial Intelligence and Real-time Systems

Background

Although the advent of self-driving cars has received the most public attention, autonomous systems are fast becoming prevalent in a wide variety of use-cases. One use-case with potentially far-reaching implications is the automation of railway traffic. Considerable progress has been made recently in the areas of object detection and scene understanding, thus improving the reliability of autonomous systems [6, 9, 4]. However, full automation has to be predicated on the provable safety and robsutness of detection algroithms responsible for recognizing and classifying scenarios during opeartion, as even small misclassifications can result in hazardous behavior.

A key challenge is the susceptibility of neural networks to adversarial attacks, sensor noise, and other disturbances [3]. Such perturbations, though small, can profoundly influence model outputs and consequently cause hazardous system behaviour with, as in the case of autonomous trains, potentially catastrophic consequences. Furthermore, verifying image-based neural networks presents a unique challenge due to the high-dimensional continuous input space [5].

Formal verification methods, which provide mathematical proofs of correct behavior, offer a promising avenue for certifying safety [8, 2]. Recent research on set-based verification, such as using zonotopes and abstract interpretation techniques, has shown promise in analyzing deep neural networks in a scalable way [7]. However, existing approaches have not yet been extended to tackle the verification of networks for entire video sequences. Additionally, they usually only analyze the robustness of neural networks in a local neighborhood around some input image. This project investigates a novel framework to apply such formal methods to networks processing video data, with the goal of identifying and certifying entire regions of the image input space where the neural network's output remains trustworthy.

Description

This project aims to demonstrate provable safety for deep neural networks used for interpreting camera footage in railway applications. It focuses on identifying regions of the input space—images or video frames—where the network's prediction can be formally proven to match expected safety outcomes.

To achieve this, the input space must be partitioned into verifiable subregions for which safety can be shown. This can be attempted via *bottom-up* clustering of similar images into such subregions, or via *top-down* specification-driven input space refinement, which attempts to translate safety specifications from the output space to the input space. The goal is to maximize the volume of these partitions while ensuring their verifiability.

By focusing on the structure of the input data and partitioning the input space into manageable pieces, we can create a system that, during runtime, quickly checks if a given image falls within a verified-safe region. If it does, the system can rely on the network's output; if not, it may raise a warning or use a fail-safe fallback mechanism.

Tasks

- Literature research of relevant work related to the task of image/video verification for neural networks.
- Familiarize with relevant aspects of the toolbox CORA [1] relating to neural network verification and specification driven input refinement.
- Implementation of the partitioning of the input space of possible images into verifiable subregions through both bottom-up image clustering as well as top-down specification-driven input space refinement.

Supervisor:
Prof. Dr.-Ing. Matthias Althoff

Advisor:
Tobias Ladner, M.Sc.

Research project:
FAI

Type:
GR

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:
14. April 2025

For more information please contact us:

Phone: +49 (89) 289 - 18140

E-Mail: tobias.ladner@tum.de

Website: ce.cit.tum.de/cps/

- Evaluate the performance of both approaches to input space partitioning.
- Optional: include a robustness heuristic for the verified safety/unsafety of images, and investigate possibilities to reduce the dimensionality of the input space.

References

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

[2] Christopher Brix, Stanley Bak, Taylor T Johnson, and Haoze Wu. The fifth international verification of neural networks competition (vnn-comp 2024): Summary and results. *arXiv preprint arXiv:2412.19985*, 2024.

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

[4] Quoc-Vinh Lai-Dang. A survey of vision transformers in autonomous driving: Current trends and future directions. *arXiv preprint arxiv:2403.07542*, 2024.

[5] Aditya Parameshwaran and Yue Wang. Scalable and interpretable verification of image-based neural network controllers for autonomous vehicles. *arXiv preprint arxiv:2501.14009*, 2025.

[6] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[7] Gagandeep Singh, Timon Gehr, Matthew Mirman, Markus Püschel, and Martin Vechev. Fast and effective robustness certification. *Advances in Neural Information Processing Systems*, 31, 2018.

[8] Caterina Urban and Antoine Miné. A review of formal methods applied to machine learning. *arXiv preprint arXiv:2104.02466*, 2021.

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Technical University of Munich

Department of Informatics

Chair of Robotics, Artificial Intelligence and Real-time Systems