Runtime Verification of Computer Vision Deep Neural Networks against Symbolic Constraints

Degree: Bachelor/Master
Supervisors: Prof. Dr. Nadja Klein (KIT), Moussa Kassem Sbeyti (KIT), Dr. Gesina Schwalbe (University of Lübeck)
Institution: KIT (Karlsruhe Institute of Technology) in collaboration with University of Lübeck
Status: Available

Abstract

Recent work has introduced simple methods to evaluate compliance with symbolic rules in black-box deep neural networks (DNNs). This thesis investigates and quantitatively compares the extent to which different computer vision DNNs comply with symbolic rules using real-world datasets, such as those from the automated driving domain. The goal is to assess how effectively existing verification and testing techniques for DNNs can identify remaining issues in the model's learned knowledge. Finally, the approach will be evaluated as a runtime verification framework for DNNs, which can be installed post hoc and trigger an alert in case of implausible outputs.

Problem Statement

Deep neural networks excel in computer vision tasks, but are yet too unreliable for use in safety-critical applications such autonomous driving. A core reason are unavoidable, but unintuitive and wrong correlations in the training data. These are easily incorporated by the DNN during training and may lead to failures in rare situations (e.g., high occlusion).

This makes it important to ensure that DNNs comply with given intuition in form of symbolic constraints on the desired outputs, for example "If there is a head, there should usually be a person" (isHead(region) => isPerson(region)). Techniques from concept-based explainable artificial intelligence (C-XAI; Lee et al. 2025) allow to associate symbols (concepts) of interest, e.g., "head", with regions in the internal representations of a trained deep neural network. As proposed by Schwalbe et al. (2022), this can be used to attach additional segmentation outputs for those concepts to a trained DNN, even if the DNN has not directly been trained to output these symbols. Subsequently, it can be tested on a test set or even during runtime, whether the extended DNN outputs fulfill the (potentially fuzzy) logical constraints.

However, so far this approach has only been showcased on a very small setup. Therefore, it remains open, how effectively the method uncovers errors for different DNN architectures, datasets, and rule sets. In other words: How many DNN failures arise from inconsistencies with known rules? And which factors influence how strongly non-compliance with logical constraints correlates with incorrect predictions?

Goals

1. Define a knowledge base of diverse rules applicable to vision tasks which serve as a test setup to test rule compliance
2. Implement the verification testing setup for a selection of concurrent vision DNN architectures, rules, and datasets
3. Conduct and evaluate a comparative study:
    - Assess what are influence factors in DNN architecture, rule type and dataset for rule compliance
    - Correlate rule compliance against quality as runtime monitor, i.e., ratio of false alarms against uncovered true errors

Approach

The comparison will follow these steps:

Architecture Comparison: How do different object detection DNN architectures compare (Convolutional DNNs / Vision Transformers; small / big models)?

Dataset Comparison: How do different datasets compare (general like MS COCO vs. automotive like A2D2)?

Setup for Symbol and Relation Extraction:

- **Rule base:** Start with a semantically rich domain like automated driving, for which plenty of intuitive rules and full ontologies are available (Giunchiglia et al. 2022)
- **Base method:** Use the C-XAI method described in Schwalbe et al. (2022), for which a rich code base is available
- **Symbols:** Simple object classes (e.g., street light, car, person), object parts (e.g., head, arm, steering wheel), and/or object attributes (e.g., red) using existing datasets like ImageNet, German Traffic Sign Datasets, or BRODEN dataset
- **Relations:** Simple hierarchical relations (isA), and 2D spatial relations (isPartOf) estimated from concept segmentations, potentially extended to 3D spatial relations using predicted or ground truth depth information
- **Logic:** Boolean and probabilistic t-norm fuzzy logic as starting point, later extended/compared against other continuous t-norm fuzzy logics (for a brief introduction see, e.g., Schwalbe et al. (2022) or Roychowdhury et al. (2018))

Requirements

- Solid programming skills in Python and familiarity with the PyTorch deep learning framework

- Familiarity with machine learning using DNNs and logistic regression models
- Familiarity with formalization of knowledge as logical rules
- Basic understanding of continuous fuzzy (multi-valued) logics

Literature

- Giunchiglia, Eleonora, Mihaela Stoian, Salman Khan, Fabio Cuzzolin, and Thomas Lukasiewicz. 2022. “ROAD-R: The Autonomous Driving Dataset with Logical Requirements.” In IJCLR 2022 Workshops. Vienna, Austria. https://arxiv.org/abs/2210.01597.
- Lee, Jae Hee, Georgii Mikriukov, Gesina Schwalbe, Stefan Wermter, and Diedrich Wolter. 2025. “Concept-Based Explanations in Computer Vision: Where Are We and Where Could We Go?” In Computer Vision – ECCV 2024 Workshops, edited by Alessio Del Bue, Cristian Canton, Jordi Pont-Tuset, and Tatiana Tommasi, 266–87. Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-92648-8_17.
- Roychowdhury, Soumali, Michelangelo Diligenti, and Marco Gori. 2018. “Image Classification Using Deep Learning and Prior Knowledge.” In Workshops of the 32nd AAAI Conf. Artificial Intelligence, WS-18:336–43. AAAI Workshops. New Orleans, Louisiana, USA: AAAI Press. https://aaai.org/ocs/index.php/WS/AAAIW18/paper/view/16575.
- Schwalbe, Gesina, Christian Wirth, and Ute Schmid. 2022. “Enabling Verification of Deep Neural Networks in Perception Tasks Using Fuzzy Logic and Concept Embeddings.” arXiv. https://doi.org/10.48550/arXiv.2201.00572. (Preprint)

Contact: For inquiries, please contact Prof. Dr. Nadja Klein (nadja.klein@kit.edu), Moussa Kassem Sbeyti (moussa.sbeyti@kit.edu), or Dr. Gesina Schwalbe (gesina.schwalbe@uni-luebeck.de).