Trustworthy Vision for Autonomous Vehicles

Zahra Chaghazardi, Saber Fallah and Alireza Tamaddoni-Nezhad

Department of Computer Science, University of Surrey
z.chaghazardi@surrey.ac.uk, s.fallah@surrey.ac.uk, a.tamaddoni-nezhad@surrey.ac.uk

💡 Project Insights 💡

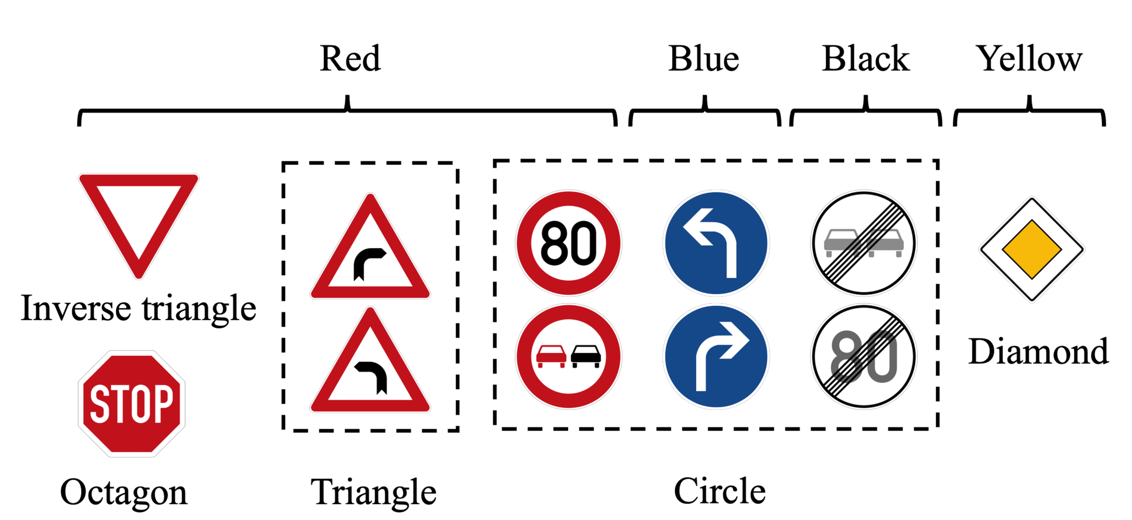
- Developed the Robust Logic-infused Deep Learning (RLDL) approach, integrating Inductive Logic Programming (ILP) with neural networks for enhanced traffic sign recognition.
- Demonstrated that incorporating logical consistency constraints improves model robustness, especially under adversarial attacks.
- Achieved significant accuracy improvements in recognizing traffic signs, contributing to safer autonomous vehicle (AVs) operations.

Introduction

→ Deep learning models, while powerful, are vulnerable to adversarial attacks, making them less reliable in safety-critical applications such as AVs.

→ To address these concerns, this project focuses on improving the reliability of deep learning models in AVs through traffic sign recognition, a crucial task for AV safety.

→ Incorporating logical consistency into deep learning models for vision tasks can improve the accuracy and reliability of AVs. Logic-based methods ensure models adhere to predefined rules, enhancing their resistance to adversarial attacks and overall reliability for AVs.



⇒ **Traffic signs and their labels.**

We trained a multi-label classifier on the GTSRB dataset [1], which includes 11 types of traffic signs.

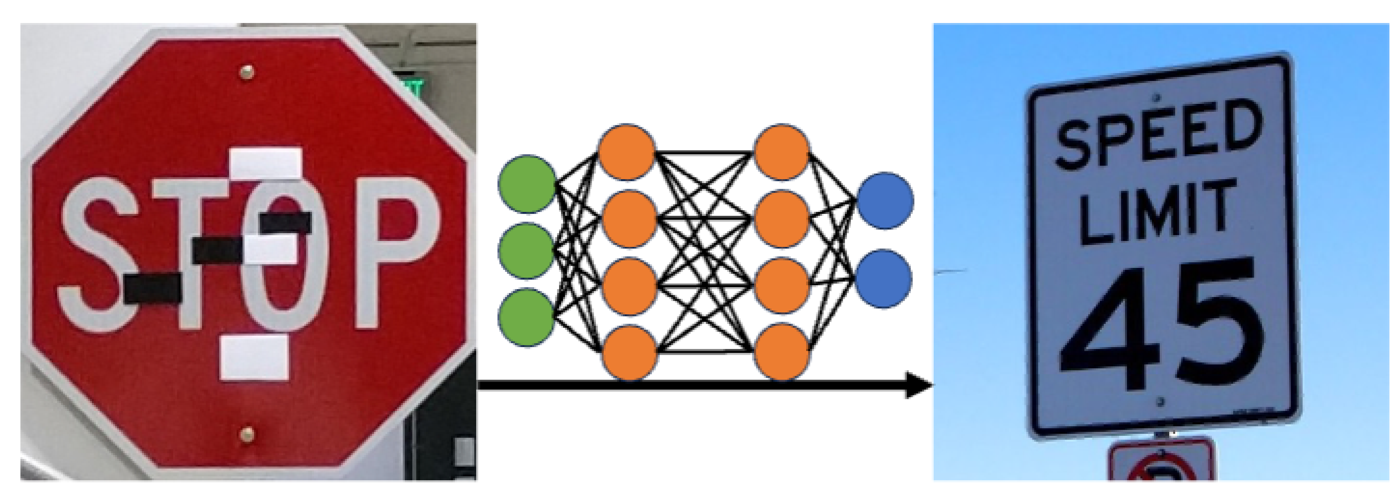
Prob Statement

A key challenge in using deep learning for autonomous vehicles is the vulnerability of neural networks to adversarial attacks. These attacks subtly modify input data, causing misclassification and endangering AVs. Despite high accuracy under normal conditions, current models struggle to perform reliably in adversarial scenarios.

Existing Research

Some existing research has explored improving deep learning robustness through adversarial training, where models are trained on adversarially modified datasets to improve resilience. Others investigate neuro-symbolic methods, which integrate symbolic reasoning with neural networks to enforce logical constraints during training and inference.

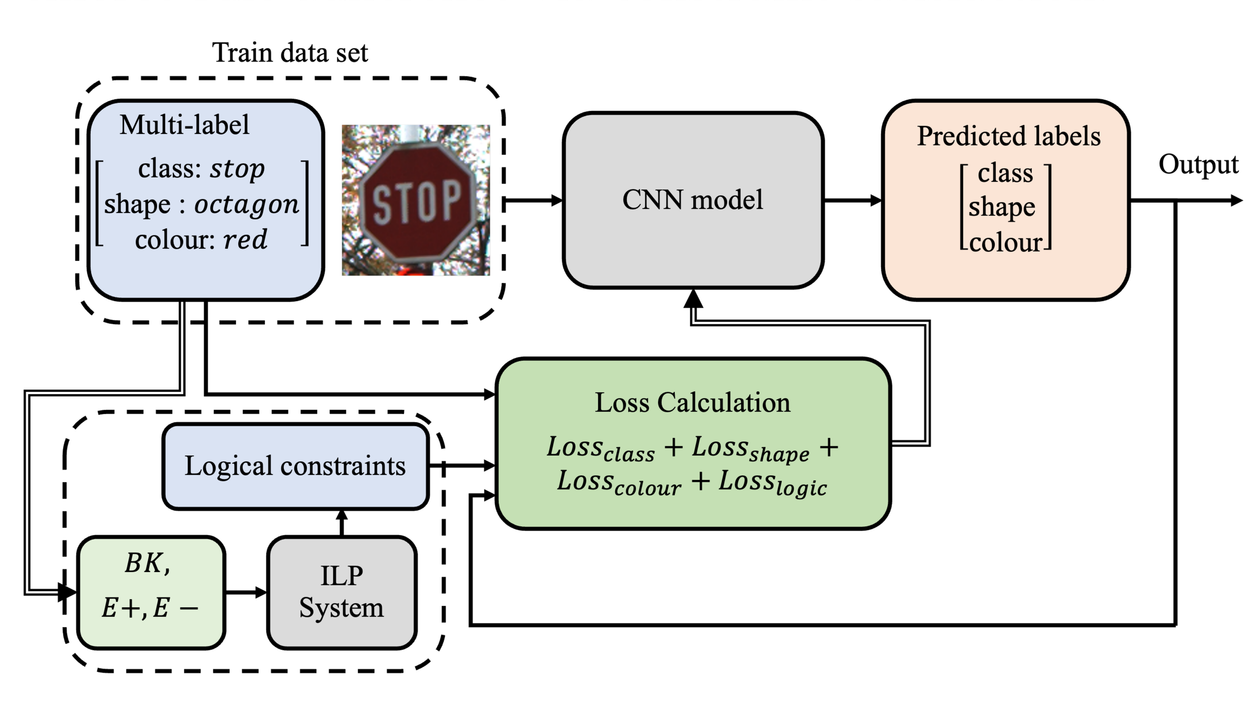
The Threat of Adversarial Attacks



⇒ **Deep Learning Vulnerability**

Methodology

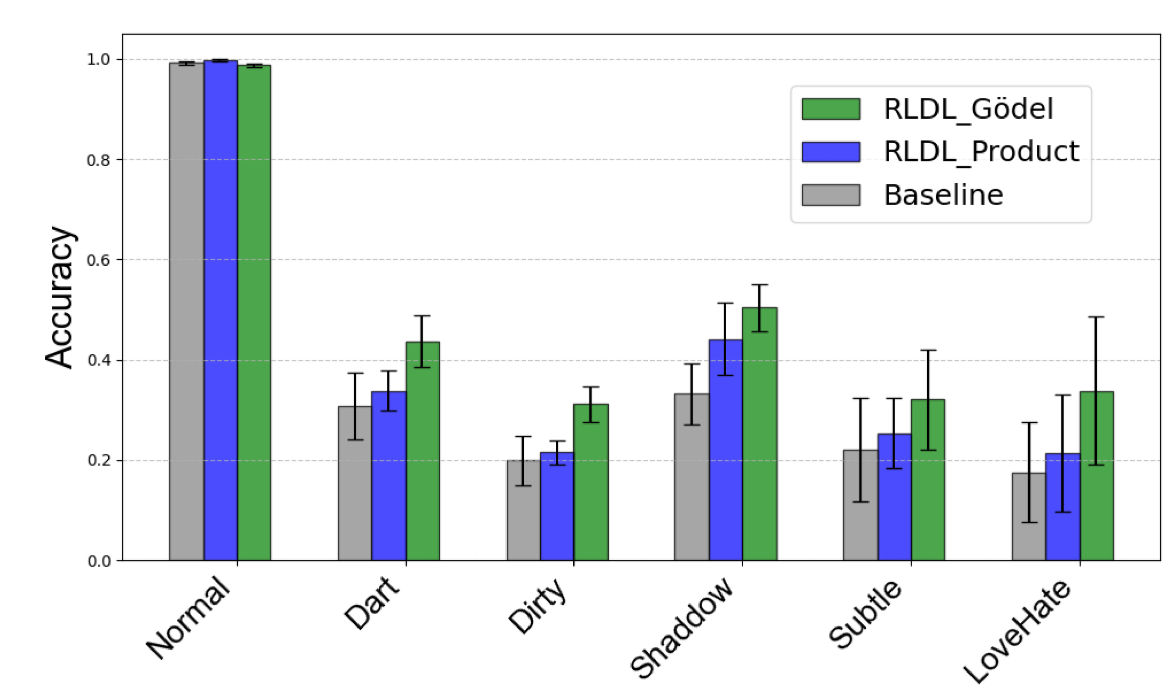
We propose a Neuro-symbolic traffic sign classifier that integrates symbolic reasoning with deep learning. A multi-label CNN model predicts traffic sign classes, shapes, and colours. During training, predictions are assessed against logical constraints, with a regularisation term added to the loss function to enforce rule adherence via constraint satisfaction.



⇒ **RLDL traffic sign classifier**

Results

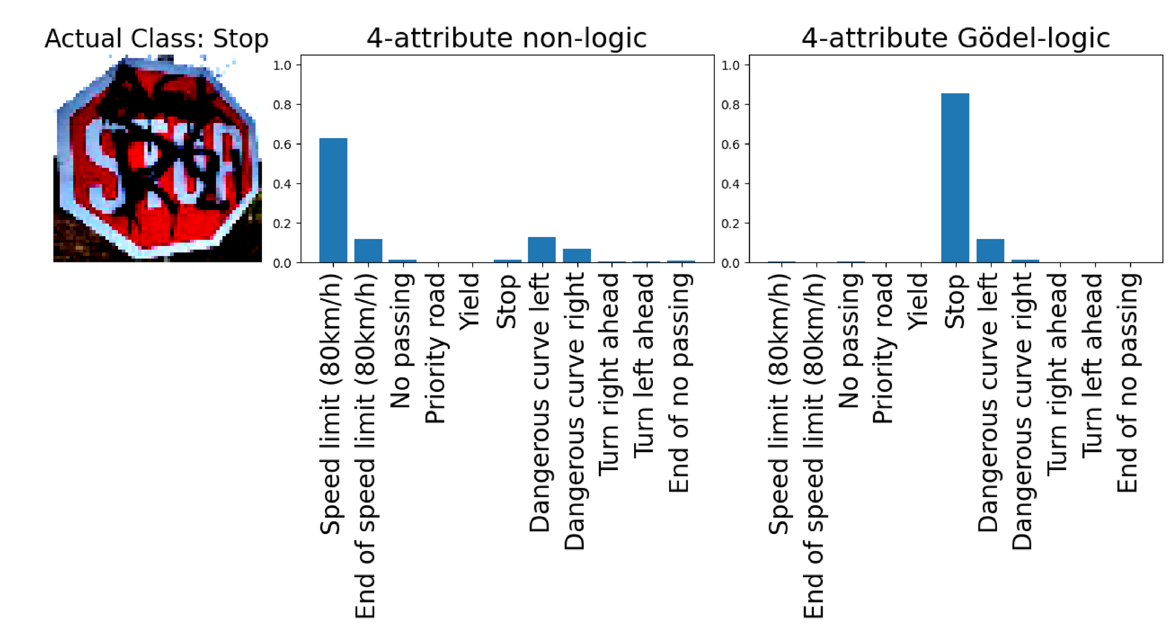
The accuracy of the baseline and proposed logic-based models on targeted stop signs has been evaluated across various datasets (Normal, Dart, Dirty, Shadow, Subtle, and LoveHate). RLDL Gödel significantly outperforms the baseline model on adversarial datasets, highlighting its superior robustness.



⇒ **Models accuracy comparison**

Conclusion

RLDL integrates logical rules, derived from Inductive Logic Programming (ILP), into the CNN model to enforce logical consistency in predictions.
Experimental results show that RLDL significantly improves the robustness of neural networks in AVs under adversarial conditions.



⇒ **Models accuracy comparison**

References 📜

[1] Stallkamp et al.
Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition.
Neural networks, 32:323–332, 2012.

Project Git Repo

UNIVERSITY OF SURREY