150 Jahre
culture of
excellence

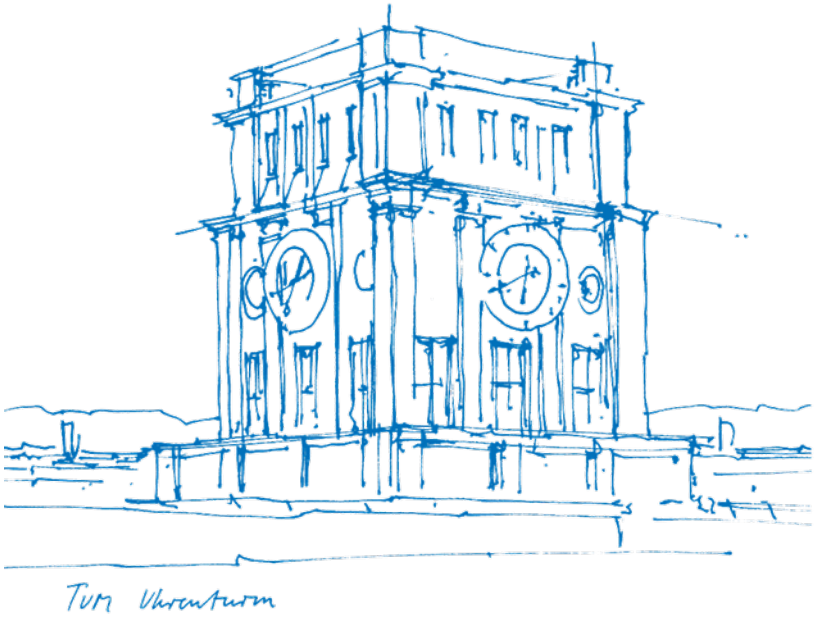
Uncertainty in Machine Learning:
A Safety Perspective on Autonomous Driving

Sina Shafaei, Stefan Kugele, Mohd Hafeez Osman and Alois Knoll

Technical University of Munich
Department of Informatics
sina.shafaei@tum.de

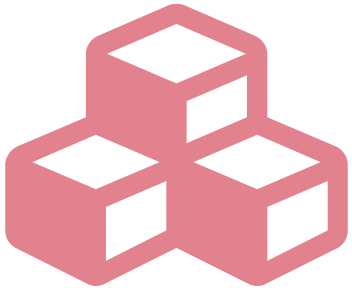
International Workshop on Artificial Intelligence Safety Engineering
WAISE 2018

Västerås, Sep. 18nd 2018



Outline

Introduction



Challenges

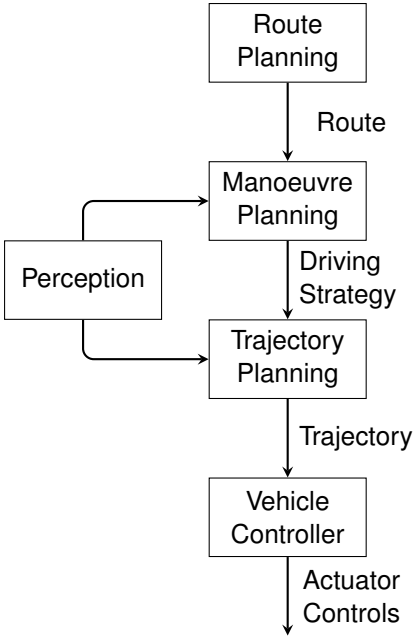
Proposed Approaches

Conclusion & Future Works

Introduction

Safety of AI Applications

- Two perspectives of *Run-time* and *Design-time*
- Serious lack of concrete approaches to address the challenges in practical manner,
- We consider a Maneuver Planning Systems as our main use case in an autonomous car,



Introduction

Safety Critical Situations – Question 1

- System is trained and tested on data from roads in a environment with well-behaved traffic
  - Deployed in a different environment with chaotic driving conditions

- System is trained and tested on roads with 4 wide lane
  - System is placed on a 2 way narrow lane road

Will this lead to an un-safe situation?

Introduction

Safety Critical Situations – Question 2

- The vehicle wants to overtake another vehicle in front,
- Based on the country, driving rules state that one must overtake only from one side (left or right),
- This is imbibed in us, humans,

What about an autonomous vehicle employing such system?

Introduction

Safety Critical Situations – Question 3

- The vehicle needs to execute a lane change operation to reach its goal state,
- At the same time, there is a vehicle on the left increasing the possibility of an accident,
- Standard deep learning techniques generate the output only as hard classifications,

May the chance of a condition with such low probability get ignored and lead to costly collisions/accidents?

Introduction

Safety Critical Situations – Question 4

- Humans are designed to be innately optimistic,
- Neural networks are normally trained to exhibit the positive outputs that we expect to receive from them,

Is there any benefit that could be reaped by getting trained to generate positive as well as negative outputs?

Challenges

Uncertainty in machine learning can be categorized into:

- *Aleatoric* (data dependent),
  - The noise in data is captured by the model, resulting in the ambiguity of training input,
- *Epistemic* (model dependent)
  - A measure of familiarity, as it represents the ambiguity, the model exhibits when dealing with operational inputs

The major causes of concern while dealing with ML-based application are as follows:

- Incompleteness of training data,
- Distributional shift,
- Differences between training and operational environment,
- Uncertainty in prediction

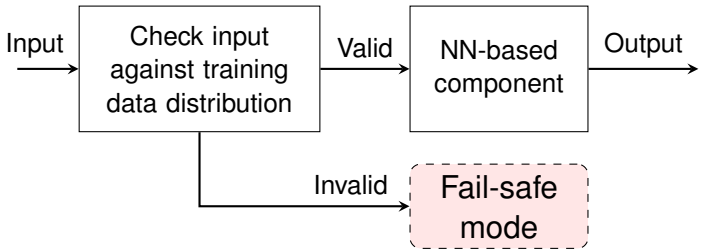
Proposed Approaches

We assume that the action to be taken in the fail-safe mode is known beforehand

Variational Methods to Filter ‘Anomalous’ Operational Inputs

Case 1

- The differences in training and operational conditions,
- How ‘far away’ is the input from the training data,
- Detecting ‘anomaly’ in input data in online manner,



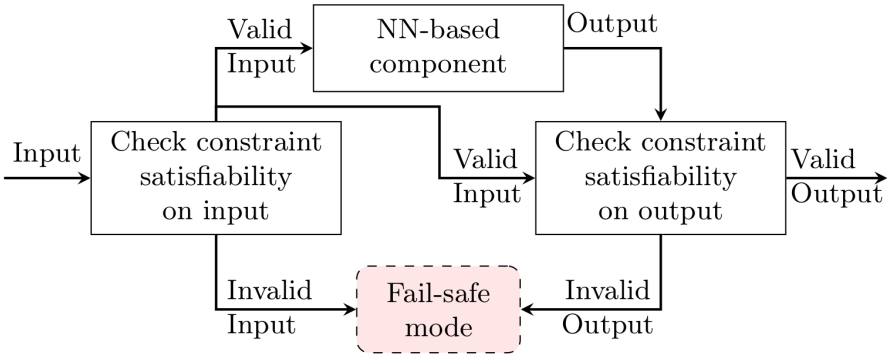
Advantage of this approach:

- Characteristics of input data learned from the data,
- No special feature engineering effort is required,
- Highly generalizable approach,
- Simply exposing the system to the data for modeling the environment, can help the system draw required inferences

Defining Environmental Constraints

Case 2 – **Consideration**: action to be taken in the fail-safe mode is known beforehand / Design-time

- A promising way to model the entities and relations,
- Ontology model is defined by the safety engineer,
- The main topics of functional safety can be derived from ISO 26262,
- The stored concepts will be translated into machine-readable first-order logic (e.g. Prolog code),
- Ontologies can be seen akin to a ‘Safety Blanket’ around each ML-based component,
- Inputs and outputs will be tested against the set of environmental constraints

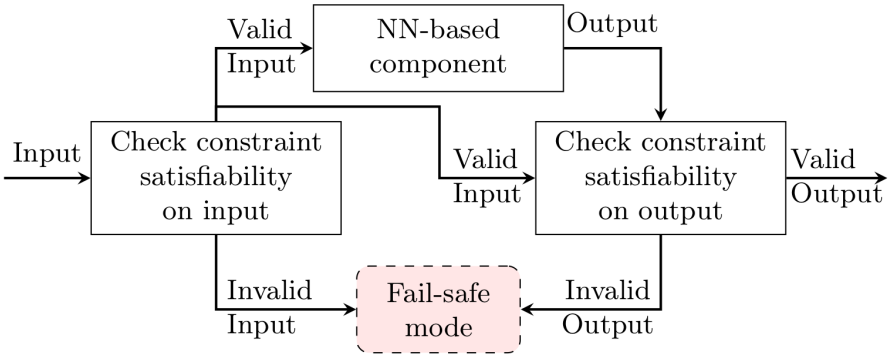


Defining Environmental Constraints

Case 2

Advantages of this approach:

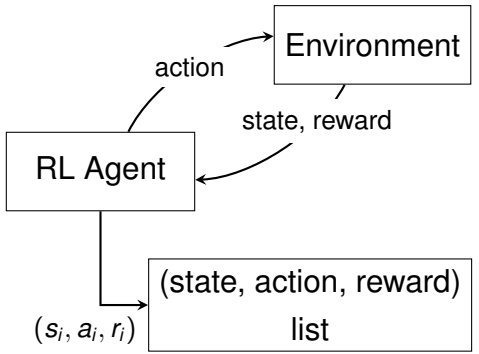
- Improved readability of the system
- Following the principles of traditional verification and validation methods,
- Ensuring the developed system abides by the intuition of human actors,
- Improvements on traceability of issues and tracking shortcomings with the system



Pre-exploration Using Reinforcement Learning

Case 3

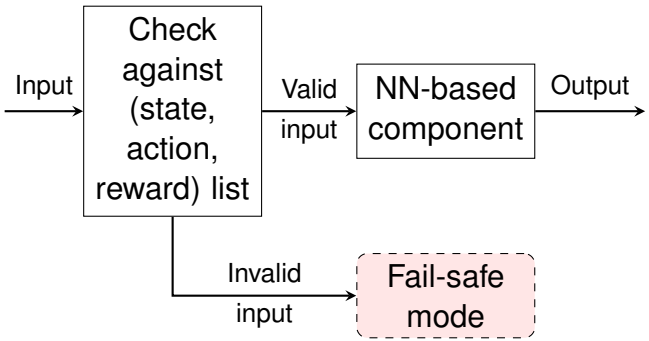
- Modeling the situation in terms of rewarding and penalizing behavior,
- Augment learning with two trainable components,
- Exploring and interacting with the environment via simulations to cover even negative outcomes,
- Generating the map of situations, actions and associated reward values,



Pre-exploration Using Reinforcement Learning

Case 3 – Cont.

- Mapping will be used for categorizing the situations that lead to a high, medium and low risk based on reward values,
- New inputs will be checked against the safety invariance mapping
- Generalizing to similar use-cases with limiting factor of additional hyper-parameter for tuning the agent

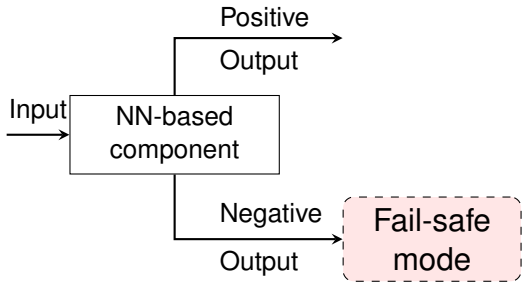


Ensuring Coverage of Positive and Negative Cases

Case 4

In the example of manoeuvre planning system, the component should be able to predict not only left, straight, and right actions, but also outputs that could lead to negative outcomes such as driving off the road, a crash and so on.

- Higher assurance of the system being trained on under-represented or rare situations/inputs,
- Well generalization of the system,
- Increasing the understandibility of behavior of the system,
- Easy to implement,



Conclusion & Future Work

Overview

1. Uncertainty is one of the main challenges of ensuring safety in autonomous driving,
2. Our use cases for different situations were just tip of an iceberg,
3. There is a lack of concrete mechanisms for ensuring the safety at Design-time and Run-time,
4. One individual technique is not enough for verifying the safe functionality of an adaptive system,
5. The focus should be on building a toolbox of different verification and validation techniques,
6. A layered approach where each layer of monitoring for data and application, independent of the other, focuses on one aspect of the safety requirement would be beneficial

Thank You.