Uncertainty in Machine Learning: A Safety Perspective on Autonomous Driving

Sina Shafaei(✉), Stefan Kugele, Mohd Hafeez Osman, and Alois Knoll

Technical University of Munich, Munich, Germany
{sina.shafaei,stefan.kugele,hafeez.osman}@tum.de,
knoll@in.tum.de

Abstract. With recent efforts to make vehicles intelligent, solutions based on machine learning have been accepted to the ecosystem. These systems in the automotive domain are growing fast, speeding up the promising future of highly and fully automated driving, and respectively, raising new challenges regarding safety assurance approaches. Uncertainty in data and the machine learning methods is a key point to investigate one of the main origins of safety-related concerns. In this work, we inspect this issue in the domain of autonomous driving with an emphasis on four safety-related cases, then introduce our proposals to address the challenges and mitigate them. The core of our approach is on introducing monitoring limiters during development time of such intelligent systems.

Keywords: Artificial intelligence · Uncertainty · Safety

1 Introduction

The safety aspect of the artificial intelligence-based applications has captured the attention of researchers recently, especially for the case of machine learning-based approaches such as neural networks and deep learning methods [1,3,9,10,14] and investigated from two different perspectives: (i) *Run-time* [7] and (ii) *Design-time* [5]. However, there is still a serious lack of concrete approaches which address the challenges in a practically efficient manner. In this work, we focus on the *uncertainty* issue of machine learning algorithms. We intuitively categorise the safety-critical situations originated from this issue, that a manoeuvre planning system may face, into four different cases. Finally, we propose approaches in order to address the challenges in each case. As mentioned, we are concentrating on the following cases in a manoeuvre planning system:

Case 1. The system has been trained and tested on the data from roads in a country with well-behaved traffic but is instead deployed for driving on roads in another country with chaotic driving conditions. Another similar case is when the vehicle has been trained and tested on roads with 4 wide lane driving but is instead faced with a 2-way narrow lane drive. In such situations, the outputs of

B. Gallina et al. (Eds.): SAFECOMP 2018 Workshops, LNCS 11094, pp. 458–464, 2018.
https://doi.org/10.1007/978-3-319-99229-7_39

the intelligent vehicle cannot be relied upon, as there is no guarantee that the system would behave as expected.

Case 2. The vehicle which employs this system wants to overtake another vehicle in front of it. Based on the country, driving rules state that one must overtake only from one side (left or right). Though this is imbibed in us, humans, while learning to drive, when it comes to autonomous vehicles there is no guarantee that the system has indeed learned this rule and will always follow it.

Case 3. The vehicle needs to execute a lane change operation to reach its goal state, but there happens to be a vehicle on the left that is in such an alignment with the ego vehicle that, though not very high probability, there is a possibility of an accident. Since standard deep learning techniques generate as output only hard classifications, there is still the chance of a condition with such low probability getting ignored and lead to costly collisions/accidents.

Case 4. Humans are designed to be innately optimistic, which might even be reflected in the training data for neural networks. NNs in autonomous vehicles are usually trained to exhibit the positive outputs that we expect to receive from them, however that benefits could be reaped by getting trained to generate positive as well as negative outputs.

2 Machine Learning and Safety Challenges

The *uncertainty* in machine learning algorithms can be categorised into two types [8]: (a) *aleatoric* or data dependent, where the noise in the data is captured by the model, resulting in the ambiguity of training input and (b) *epistemic* or model dependent seen as a measure of familiarity, as it represents the ambiguity the model exhibits when dealing with operational inputs. More precisely the major causes of concern while dealing with ML-based solutions are as follows:

(i) ***Incompleteness of Training Data*** – Traditional software systems are developed with a pre-defined set of functional requirements. However, in NNs, and more generally in ML algorithms, the functional requirements of the system are implicitly encoded in the data that it is trained on, expecting that the training data represents the operational environment. The setback, however, is that training data is by definition incomplete [11], as it represents a subset of all possible inputs that the system could encounter during operation. Insufficiencies thus arise when the operational environment is not wholly represented in the training set. In the case of autonomous vehicles, critical and ambiguous conditions, where the vehicle is expected to act predictably, usually tend to be problematic. This is because such situations, owing to their either extremely rare or highly dangerous nature, tend to be underrepresented in the training set [1,3].

(ii) ***Distributional Shift*** – In the case of an autonomous vehicle, the operational environment is highly unpredictable [3] as it is constantly changing in response to the actors within the system. Therefore, even with a good and near perfect training set, the operational inputs may not be similar to the training set. In other words, there could be a shift in the distribution of operational

data as compared to the original training data, resulting in the system behaving unpredictably.

(iii) ***Differences between Training and Operational Environments*** – Subtle changes in the operational environment can lead to a state of unpredictable behaviour [3] in NNs. An NN fine-tuned for a certain specific setting provides no guarantee of working in the exact same way when the settings are changed.

(iv) ***Uncertainty of Prediction*** – Every NN has an error rate associated with it [11], and the aim of the training process is to reduce this error rate as much as possible. In the operational environment, this error rate can be interpreted as an *uncertainty* associated with the output produced by the model. Though this *uncertainty* can tell us about how well the system models the environment, it is not accounted for in the cyber-physical systems of today [8].

3 Proposed Approaches

Due to the fact that we are not able to handle all of the safety-critical situations, in our proposed approaches, we assumed that the action to be taken in the fail-safe mode is known beforehand and could include actions such as slowing down the car, bringing the car to a halt, or even handing over control to the human driver. Moreover, since we are focusing on safety for any AI-related software, the risk assessment is not in the scope of this paper.

3.1 Variational Methods to Filter 'Anomalous' Operational Inputs

(Case 1.) This method targets the problems related to differences in training and operational conditions and builds on the idea of online data monitoring. The main intuition (as depicted in Fig. 1) is to detect how 'far away' is the input from the data the system was trained on. In other words, the aim is to detect if the input is an 'anomaly', i. e., a data point that is significantly different from the original data. If yes, then the system is expected to enter a fail-safe mode, else normal operation continues. Given some data X, Variational Inference (VI) [2] aims to find a distribution $\mathrm{Q}(Z)$ which is as similar to the true posterior $\Pr(Z \mid X)$ as possible, where the distance between the distributions can be calculated using the *Kullback-Liebler Divergence* a.k.a. relative entropy. Use of variational inference [2] is proposed for this online detection of anomalies [12]. The advantage of this approach is that the characteristics of expected input are learned from the data, and so no special feature engineering

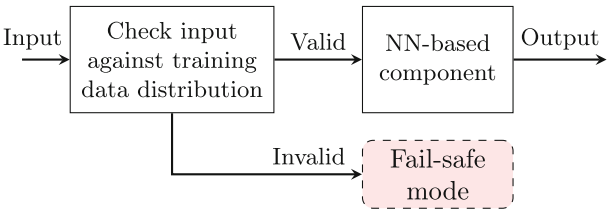

Fig. 1. Control flow of anomaly detection approach

efforts are required. This also means that this approach is highly generalisable and is not bound by the use case. Simply exposing the system to data for modelling the environment, can help the system draw required inferences.

3.2 Defining Environmental Constraints

(Case 2.) We propose the use of ontologies to enforce such conditions as depicted in Fig. 2. Ontologies are a way to model the entities and relations in a system [4]. During *design-time*, the automotive safety engineer needs to create an automotive safety ontology (based on specific software/system function and context). The main ontology topics (for functional safety) can be derived from ISO 26262 (Part 1 - Vocabulary) [6]. The concepts stored in ontologies will be internally translated into machine-readable first-order logic (e.g. Prolog code), thereby making it simpler for describing constraints that the system must obey in the environment. Ontologies can be seen akin to a 'safety blanket' around each ML-based component. Inputs to the component and outputs generated thereby will be tested against the set of environmental constraints to ensure that they are fulfilled, if not, the system enters a fail-safe mode. This solution improves the reliability of the system, and follows the principles of traditional verification and validation methods, ensuring that the developed system abides by the intuition of human actors. It can improve traceability of issues and can also help track shortcomings with the system.

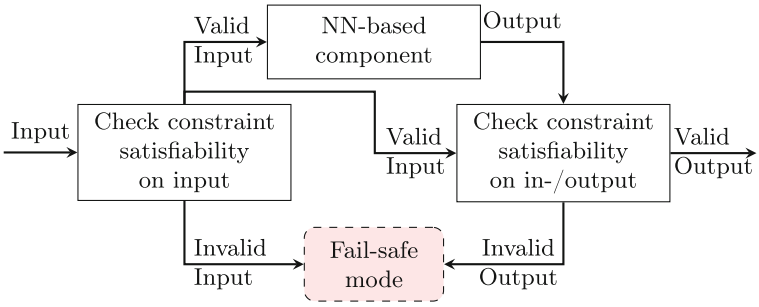

Fig. 2. Control flow of ontology-based constraint satisfaction approach

3.3 Pre-exploration Using Reinforcement Learning

(Case 3.) Since such situation can be modelled in terms of rewarding and penalising behaviour, we suggest the use of a reinforcement learning (RL) agent to mitigate such conditions. Reinforcement Learning [13] is based on behaviourist psychology, wherein the agent learns to map situations to actions, by receiving rewards from the environment for good behaviour and penalties otherwise. The aim for this solution is to augment learning with two trainable components, as shown in Fig. 3. Figure 3a shows the RL agent that is responsible for exploration of the environment, and Fig. 3b describes the online NN that is implemented in the standard manner for the component in question. The RL-agent learns

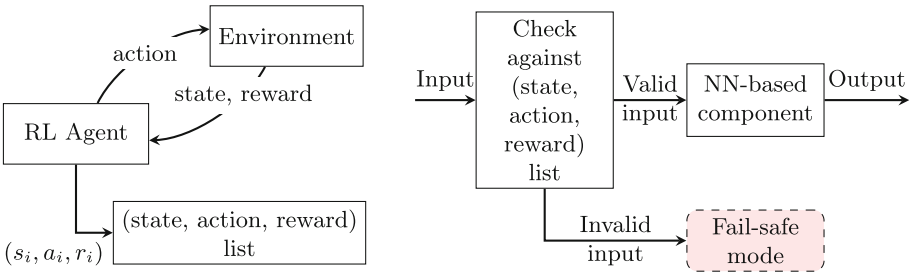

Fig. 3. Control flow of RL-based pre-exploration approach

by exploring and interacting with its environment, and so would be trained via simulations to explore even negative outcomes, as in testing these do not pose a real threat to lives. In doing so the RL-agent would be able to learn and thereby generate a map of situations, actions, and associated reward values. This mapping can then be used to categorise situations that lead to high, medium, or low risk based on the reward values of each state. This approach can be seen as an extension to the monitoring techniques, wherein, rather than manually labelling the state space as being safe or not, the output of RL agent is used to generate such a mapping, with the reward function determining the severity of the hazard for each state-action pair. Thus, every input being passed to the NN-based component would first be checked against the safety invariance mapping to enter a fail-safe mode when the input is in a catastrophic zone. When it comes to generalising to other use cases, this approach could do quite well with the limiting factor of additional hyperparameter tuning for the agent. The advantage of such an approach is that rewards and objective functions can also be set up to be more aligned with human intuition, thus making the system more compliant with human expectations.

3.4 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 lateral and longitudinal actions, but also outputs, that could lead to negative outcomes such as driving off the road, a crash and so on. In such a system, if the output of a workflow falls in a negative class, the system would enter a fail-safe mode, else, would continue functioning normally, as visualised in Fig. 4. This setup brings along the benefit of higher assurance of the system being trained on under-represented or rare situations/inputs as well, leading to a better response to safety-critical

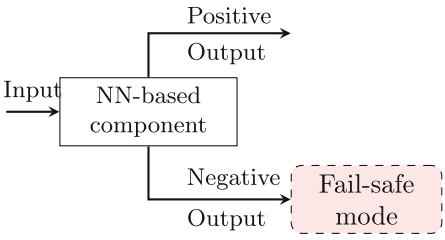

Fig. 4. The control flow of predicting possible positive and negative outputs

situations. Since the system learns expected good and bad outputs from the data directly, without explicit specifications, the system would generalise well to other use cases, too. It also has the advantage of being easy to implement and understand.

4 Conclusions and Future Work

In this work, we have investigated several challenges in ensuring safety of machine learning-based methods in the autonomous driving domain. Our main focus was on uncertainty issue which is not originated only from machine learning methods but also training data. We have considered multiple highly safety-critical situations in autonomous driving which could be the result of uncertainty issue and proposed the most promising candidates for monitoring approaches in order to preserve the safety of such system. It is worth mentioning that applying just one individual technique is not enough to verify the functionality of an adaptive software, as each has its own set of pros and cons. Instead, we need to focus on building a toolbox of different verification and validation techniques that can be applied based on the specific needs and specifications of the system. We suggest the use of a layered approach where each layer of monitoring for data and application, independent of the other, focuses on one aspect of the safety requirement.

References

1. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., Mané, D.: Concrete problems in AI safety. CoRR abs/1606.06565 (2016)
2. Blei, D.M., Kucukelbir, A., McAuliffe, J.D.: Variational inference: a review for statisticians. J. Am. Stat. Assoc. **112**(518), 859–877 (2017). https://doi.org/10.1080/01621459.2017.1285773
3. Burton, S., Gauerhof, L., Heinzemann, C.: Making the case for safety of machine learning in highly automated driving. In: Tonetta, S., Schoitsch, E., Bitsch, F. (eds.) SAFECOMP 2017. LNCS, vol. 10489, pp. 5–16. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66284-8_1
4. Feld, M., Müller, C.: The automotive ontology: managing knowledge inside the vehicle and sharing it between cars. In: Proceedings of the 3rd International Conference on Automotive User Interfaces and Interactive Vehicular Applications, pp. 79–86. ACM (2011). https://doi.org/10.1145/2381416.2381429
5. Huang, X., Kwiatkowska, M., Wang, S., Wu, M.: Safety verification of deep neural networks. In: Majumdar, R., Kunčak, V. (eds.) CAV 2017. LNCS, vol. 10426, pp. 3–29. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-63387-9_1
6. International Organization for Standardization: ISO 26262: Road vehicles-functional safety. International Standard ISO/FDIS 26262 (2011)
7. Kane, A., Chowdhury, O., Datta, A., Koopman, P.: A case study on runtime monitoring of an autonomous research vehicle (ARV) system. In: Bartocci, E., Majumdar, R. (eds.) RV 2015. LNCS, vol. 9333, pp. 102–117. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-23820-3_7

8. McAllister, R., et al.: Concrete problems for autonomous vehicle safety: advantages of Bayesian deep learning. In: International Joint Conferences on Artificial Intelligence, Inc. (2017). https://doi.org/10.24963/ijcai.2017/661
9. Pei, K., Cao, Y., Yang, J., Jana, S.: Towards practical verification of machine learning: The case of computer vision systems. CoRR abs/1712.01785 (2017)
10. Russell, S., Dewey, D., Tegmark, M.: Research priorities for robust and beneficial artificial intelligence. AI Mag. **36**(4), 105–114 (2015)
11. Salay, R., Queiroz, R., Czarnecki, K.: An analysis of ISO 26262: Using machine learning safely in automotive software. CoRR abs/1709.02435 (2017)
12. Sölch, M., Bayer, J., Ludersdorfer, M., van der Smagt, P.: Variational inference for on-line anomaly detection in high-dimensional time series. CoRR abs/1602.07109 (2016)
13. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. Adaptive Computation and Machine Learning. MIT Press, Cambridge (1998)
14. Taylor, J., Yudkowsky, E., LaVictoire, P., Critch, A.: Alignment for advanced machine learning systems. Machine Intelligence Research Institute (2016)