Perspectives on the Validation and Verification of Machine Learning Systems in the Context of Highly Automated Vehicles

Werner Damm
C. v. Ossietzky University
26111 Oldenburg, Germany

Martin Fränzle
C. v. Ossietzky University
26111 Oldenburg, Germany

Sebastian Gerwinn
OFFIS e. V.
Escherweg 2, 26121 Oldenburg

Paul Kröger
C. v. Ossietzky University
26111 Oldenburg, Germany

Abstract

Algorithms incorporating learned functionality play an increasingly important role for highly automated vehicles. Their impressive performance within environmental perception and other tasks central to automated driving comes at the price of a hitherto unsolved functional verification problem within safety analysis. We propose to combine statistical guarantee statements about the generalisation ability of learning algorithms with the functional architecture as well as constraints about the dynamics and ontology of the physical world, yielding an integrated formulation of the safety verification problem of functional architectures comprising artificial intelligence components. Its formulation as a probabilistic constraint system enables calculation of low risk manoeuvres. We illustrate the proposed scheme on a simple automotive scenario featuring unreliable environmental perception.

Modern AI and especially machine learning (ML) components are believed to be a key enabler for bringing highly automated driving functions at SAE levels 4 to 5 (SAE and others 2014) onto the market. Before such systems can be released, obtaining a rigorous guarantee of their safety is essential: systematic faults within the design (including the training phase of ML based algorithms) could have dramatic effects on the overall safety of the mass-marketed system implementations and hence also for their societal acceptance. A key challenge for this verification is the inherent uncertainty involved in object identification. To illustrate the impact of such uncertainties, consider the following artifical example of a misperception (see Fig. 1).

At time t_0, the EGO vehicle (E) has detected another vehicle v_1 on the left lane using information from a camera and RADAR sensors. At a later time instant t_1, the vehicle v_1 has closed the gap to EGO and consequently is detected still. Additionally, another vehicle v_2 has been detected at very short distance in front of EGO, while another detector has recognized the presence of a bridge in front. In this situation, EGO is confronted with the decision to either perform an overtaking manoeuvre – thereby risking a collision with v_1, or to perform an emergency brake to mitigate a potential collision with vehicle v_2. A third option would be to perform an evasive manoeuvre to the right, thereby risking a collision with a bridge pillar. Note that at t_0, the space in front of the EGO vehicle has been perceived as free. In this scenario, we assume that the time gap $t_1 - t_0$ is insufficient for a vehicle v_2 to be outside warning range at t_0 and to get to the position (and speed) perceived for v_2 at t_1, given the physical constraints on vehicle dynamics. Thus, the results of the different detectors evidently are contradictory.

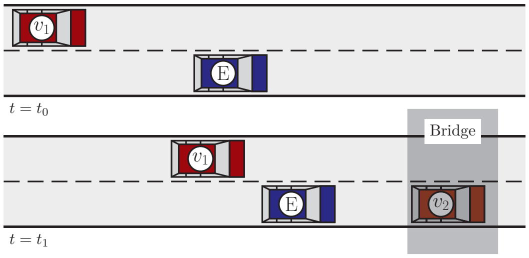

Figure 1: Example scenario. Perception of the environment is considered at two distinct time instants t_0 and t_1.

To choose an acceptable manoeuvre, a careful assessment of the risks on a vehicle level is necessary – for example by quantifying possible outcomes of a decision using injury risk scales, like AIS or ISS (MacKenzie, Shapiro, and Eastham 1985). Individual ML components, however, are traditionally evaluated using component level loss functions (Cesa-Bianchi, Conconi, and Gentile 2004). Using the common 0-1 loss ($l_{1\text{-}0}$), the resulting risk at the component level can be interpreted as bound on the probability of correctly classifying a random input (distributed according to a fixed but unknown distribution):

$$1 - \mathbb{E}[l_{1\text{-}0}] = P(\text{correctly classified}) \in [\underline{p}(\delta), \overline{p}(\delta)] \qquad (1)$$

where the right hand side denotes the confidence interval as obtained from the available bounds, i.e. via cross-validation or generalisation bounds such as within the Probably-Almost-Correct (PAC) framework. These bounds in turn depend on the confidence level δ. Under the assumption that any new data (different from the training data) would be generated according to the same probability distribution which also generated the training data, a generalisation statement can be formulated and proven which provides the desired bound on the true risk.

In order to use such information to assess the risk on vehicle level, we propose a layered approach integrating the individual ML components into a constraint system which includes prior knowledge about physical properties and the functional architecture. The resulting architecture thereby combines features from probabilistic graphical models (Koller and Friedman 2009) capturing probabilistic relationships with features from non-deterministic constraint systems. We consequently employ the same definition of risk as used in reliability and utility theory (expected loss), yet permit underspecification of the probability distribution determining the expected values of interest. Among the possible instants of the underspecified distribution, we aim at calculating worst-case expectations. This permits to compute *robust* low-risk manoeuvres at runtime, whereby individual performance assessment in terms of the empirical risk at component level can be combined with the obtained constraint system to bound the overall risk at vehicle level.

In the following, we will illustrate the proposed approach on the above example, thereby illustrating its potential.

The Probabilistic Constraint System

In the example of Fig. 1, we are interested in the following analysis questions: Can we compute a robust low-risk manoeuvre for EGO at t_1, which keeps risk adequately bounded despite potentially uncertain information? Given such a robust manoeuvre, can we quantify the worst-case residual risk associated with such controller?

To answer such questions, we first construct a constraint system reflecting assumed knowledge as well as imperfect information about the underlying situation. To this end, we try to build a probabilistic system similar to a dynamic Bayesian network (Murphy and Russell 2002). In practice, we sometimes have to admit unknown dependencies not expressible in standard Bayesian networks. For such dependencies, we possess no explicit probability distribution, but can only model constraints. We illustrate such a constraint system in Fig. 2, where the functional architecture is reflected on the left side whereas information about the real world is depicted on the right side. In the following, we refer to each signal or measurement (nodes within the figure) as variables, which can be interpreted as (possibly Dirac distributed) random variables.

We assume that EGO's sensor system provides a glare detector, a bridge detector, and a vehicle detector tracking multiple vehicles. The result of each detector is an observed variable within a Bayesian network (left side of Fig. 2). As the environment and hence also the observation thereof evolves over time, each variable is also annotated with a time index t_0, t_1 (represented as shaded duplicates of the nodes). We assume the functional architecture to be given. Hence, the Bayesian Network on the left side can be constructed with known dependencies (illustrated as thin arrows). These can contain safety mechanisms like the "Fused Vehicle Detection", which employs detection of glare to improve raw object detection by situationally reducing the importance of camera-based detection. As these are only percepts of objects, corresponding real-world counterparts are modeled on the right side. Within the dynamic Bayesian network, these counterparts act as latent variables of which dependencies and probability distributions are unknown to us. Labeled test data, however, provide values for these variables on an individual data-point basis. Physical dynamical constraints, if available, furthermore restrict their possible evolution over time. Both types of information yield an overall constraint system confining possible instantiations of the unknown distributions and thus permitting to assess worst-case (across possible instantiations) residual risk of the resulting system.

Probabilistic constraints

Using access to ground truth data from manual labeling, probabilistic constraints can be derived in terms of component based performance (Eq. 1) using standard test-scores. Within our example, the performance of vehicle detection could specify a constraint on the conditional probability

$$P\left(\widehat{v}_i \mid \text{Glare } \wedge v_i \wedge \text{Bridge}\right) \in \hat{p} \pm \epsilon(\delta) \ , \qquad (2)$$

where $\hat{p}$ denotes the empirical performance, $\epsilon(\delta)$ denotes the accuracy of such an estimate depending on the confidence level δ, and v_i denotes vehicle v_i's actual presence whereas $\widehat{v}_i$ represents that v_i was detected. Analogously, fluctuations of sensor readings can be described as probability distributions conditioned on environmental states. Although some (in-)dependence connections might be known, the explicit probability distribution might be unknown. Therefore, instead of fully specifying a dynamic belief network over all discrete and continuous variables, we only collect an incomplete set of constraints of the form of Eq. (2). This necessitates an optimisation over the possible instantiations of such underspecified distributions when calculating a safe bound on the residual risk.

Dynamic constraints

In addition to such probabilistic constraints originating from individual component tests, prior knowledge about the dynamics can be incorporated (blue box 'dynamic constraints' in Fig. 2). The detected positions of vehicles v_1 and v_2 can for example be constrained via kinematic constraints of the vehicles. Such constraints can be represented as follows, where $\ell_i(t)$ denotes the position of vehicle i at time t and $\overline{v}, \overline{a}$ are intervals containing minimal and maximal values for velocity and acceleration:

$$\ell_i(t+\Delta t) \in \left(\ell_i(t) + (\Delta t\overline{v} + \frac{1}{2}\overline{a}(\Delta t)^2)\right) \qquad (3)$$

Additional ontological constraints can reflect prior knowledge about the allowed relationship of detected objects.

As we have thus formalised a system involving variables on vehicle level ϕ as well as corresponding variables in the real world ψ, we can now relate systemic, real-world loss (e.g., in terms of available injury risk scales) to vehicle-level variables. As the vehicle variables include decision and actuator variables, such a loss function $l(\phi, \psi)$ evaluates the real-world severity of detecting, deciding, and acting. Note that both types of variables are collections of variables and in particular include references to different temporal instances.

Risk assessment

As mentioned earlier, we are interested in the overall risk of the designed function R as well as a situational risk R^s from

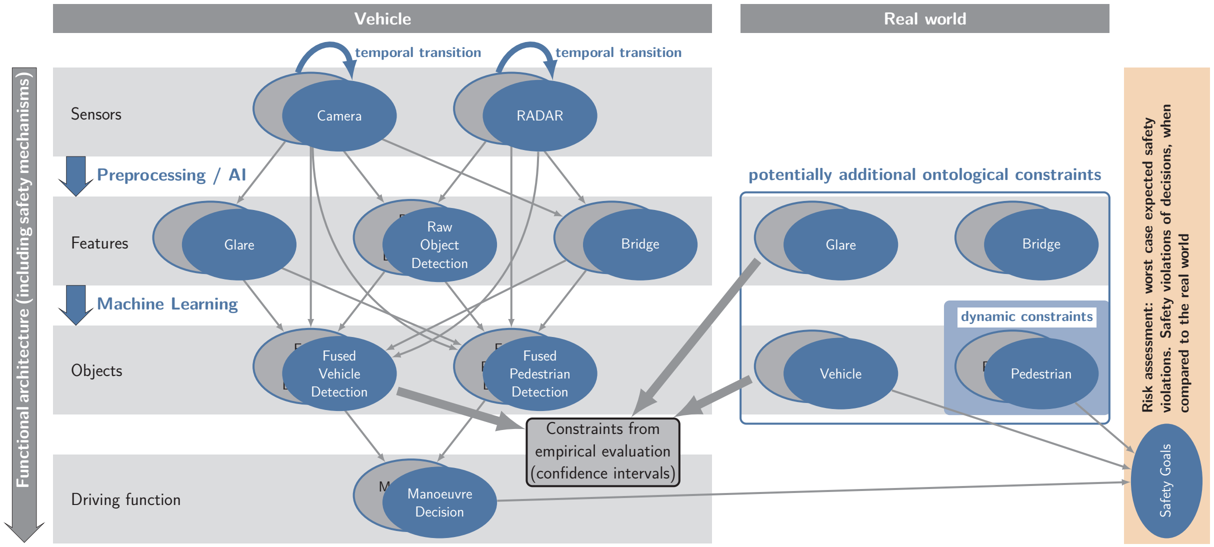
Vehicle
Real world
temporal transition
temporal transition
Functional architecture (including safety mechanisms)
Sensors
Camera
RADAR
Preprocessing / AI
Features
Glare
Raw Object Detection
Bridge
potentially additional ontological constraints
Glare
Bridge
Machine Learning
dynamic constraints
Objects
Fused Vehicle Detection
Fused Pedestrian Detection
Vehicle
Pedestrian
Constraints from empirical evaluation (confidence intervals)
Driving function
Manoeuvre Decision
Risk assessment: worst case expected safety violations. Safety violations of decisions, when compared to the real world
Safety Goals

Figure 2: Structure of the probabilistic constraint system generated from the functional architecture and the constraints obtained via empirical evaluation as well as ontological and dynamic constraints. See text for more details.

which we can derive a robust low-risk manoeuvre in a given situation. Mathematically, these quantities can be described as the following expectations:

$$R = \mathbb{E}_{(\phi,\psi)}[l(\phi,\psi)],\; R^s = \mathbb{E}_{(\psi|\phi)}[l^s(\phi,\psi)] \qquad (4)$$

Note that for the situational risk, we use the conditional distribution conditioned on observations obtained in the particular situation and a potentially different loss-function l^s (compared to the overall risk). More specifically, within the overall risk for the designed function, we might, e.g., want to use a binary loss function assigning $l(\phi,\psi) = 1$ if the situation was handled successfully and $l(\phi,\psi) = 0$ else. For the situational risk, we might want to use a quantitative assessment of the outcome. In contrast to the common setting of dynamic Bayesian Networks, the joint distribution $p_{\phi,\psi}$, however, is not completely given. Instead, only constraints over such a distribution are known due to equations like (2). More precisely, constraints as in (2) can be written as projections of the joint distribution using Bayes Rule:

$$\begin{aligned} &P\left(\widehat{v}_i \mid \text{Glare } \wedge v_i \wedge \text{Bridge}\right) \qquad (5)\\ &= \frac{P\left(\widehat{v}_i \wedge \text{Glare } \wedge v_i \wedge \text{Bridge}\right)}{P\left(\text{Glare } \wedge v_i \wedge \text{Bridge}\right)}\;, \end{aligned}$$

where each of the constraint variables either is a variable of the vehicle domain or of the real world (see Fig. 2). As the expression above omits some of the variables defined in those domains, the corresponding expressions have to be obtained by marginalising $p_{\phi,\psi}$. The question whether the (overall or situational) residual risk meets a desired bound ϑ can be formulated as a noisy optimisation problem

$$\max_{p_{\phi,\psi}\in\mathcal{P}} \mathbb{E}_{(\phi,\psi)}[l(\phi,\psi)] \overset{?}{\leq} \vartheta,\;\; \max_{p_{\phi,\psi}\in\mathcal{P}} \mathbb{E}_{(\psi|\phi)}[l^s(\phi,\psi)] \overset{?}{\leq} \vartheta\,, \qquad (6)$$

where the different constraints restrict the possible distributions, in the above formulation denoted by the set $\mathcal{P}$. If all variables are discrete, constraints on the distribution can directly be encoded into constraints on the distribution-values for different valuations of the vehicle or real-world variables. For continuous variables, the distribution has to be parametrised accordingly. Both types of constraints, however, can be incorporated into possibly non-linear functions g_i acting on the parametrised version of the distribution and the variables ϕ, ψ. For the empirical constraint of Eq. (2,5), such functions can be formalised as follows:

$$C_i(P,\phi,\psi)\text{ def.: } g_i(P,\phi,\psi) \leq c_i \qquad (7)$$

$$\underbrace{\frac{\int p(\phi,\psi)d((\phi\cup\psi)\setminus\{\widehat{v}_i, v_i,\text{ Glare, Bridge}\})}{\int p(\phi,\psi)d((\phi\cup\psi)\setminus\{v_i,\text{ Glare, Bridge}\})}}_{:=g_0(P,\phi,\psi)} \leq \underbrace{\hat{p}+\epsilon(\delta)}_{:=c_0}$$

Using specification techniques of stochastic satisfiability modulo theory (Fränzle, Hermanns, and Teige 2008), the problem (6) can alternatively be formulated as:

$$\exists_{P:\bigwedge_i C_i(P,\phi,\psi)} \rotatebox{180}{R}_{\phi,\psi\sim P} : l(\phi,\psi) \overset{?}{\leq} \vartheta \qquad (8)$$

Here, we collected all constraints over the distribution as well as over the variables within the conjunction $\bigwedge_i C_i$. Exploiting importance sampling for Eq. 8 (Fränzle et al. 2015), such problem can be made amenable for analysis using available tools (Fränzle, Gao, and Gerwinn 2017). To address scalability issues, one can also resort to statistical model checking (Ellen, Gerwinn, and Fränzle 2014).

Verification and situational analysis

Calculating the maximal risk as formalised in the previous section provides quantitative evidence to an overall safety verification process on vehicle level. Depending on the number of constraints with confidence statements, one can calculate an overall confidence level on the risk as well. Each confidence-based constraint holds with a certain confidence. If these can be regarded as independent, the overall confidence level is merely the product of the individual confidence levels. In case one is not willing to assume independence between the confidence-based constraints, the overall confidence level can be incorporated in a way similar to probabilistic constraints like (2). Note that such constraints also include constraints like c-approximate-independence as used in (Shalev-Shwartz, Shammah, and Shashua 2017), however we allow for even more pessimistic bounds whenever less information about the dependence is available.

The calculation of the maximal risk can also be performed in a particular situation. Instead of marginalising variables for the expected loss in (4), we can fix the valuation of vehicular variables to the observed values. The maximal risk then enables one to identify the most critical real-world situations and to choose a minimal risk manoeuvre. For our example, this facilitates inferring whether it is indeed more likely to falsely detect v_2 at time t_1 than having it not detected at time t_0. As due to the dynamic constraint, either v_2 has been missed at time t_0 and correctly classified at t_1 or the other way around, this restricts the joint distribution to assign zero probability to the other possibilities. Together with the empirical evidence constraints (e.g., marginal probabilities observing glare or the probability of bridges occurring), we can therefore calculate which of the two remaining possibilities are more likely. As such, it can be interpreted as the worst case interpretation of a Bayesian filter for dynamical systems which can be applied at each point in time. However, as worst-case configurations have to be identified, scalability of such an approach remains to be demonstrated in practice, but is outside of the scope of this short-paper.

Discussion

We presented a framework designed for computing (a) the current risk under given observations and (b) the overall risk under the given constraints and marginal probabilities arising from empirical evaluations of different machine learning components involved within the functional architecture.

Within our setting, such quantities are different from inference tasks typically considered within Dynamic Bayesian Networks. The central issue is that probability distributions need not completely be known, but can be underspecified, as illustrated by the occurrence of glare or bridges provide constraints on the marginal. In fact, earlier approaches in combining constraints with Bayesian Belief Networks were frequently restricted to representing constraints as pseudo-observations (Crowley, Boerlage, and Poole 2007) or to interpreting the standard inference scheme as constraint propagation (Pearl 1985). But both can also be combined to render the inference machinery more suited for such kind of constrained network (Gogate and Dechter 2012).

Automatically learning the structure of Bayesian Networks has also been explored (Berg, Järvisalo, and Malone 2014). In such an approach, constraints about the parameters (or structure) of the underlying graph can be considered. As it fits the network parameters such that the network best explains a given dataset, that approach does not immediately fit into our robust safety verification setting.

In our work, unknown or underspecified relations between variables of the network are understood as spanning and constraining a set of possible distributions. From a frequentist point of view compatible with quantitative safety, we would like to compute worst and best case scenarios under all possible assignments across the viable probability distributions rather than missing information about the dependency of different variables. This paper explains the pragmatics and the underlying mathematical constructions; the development of scalable tools automating such reasoning as well as their benchmarking remain issues of future work.

Acknowledgements

This research was partially funded by Deutsche Forschungsgemeinschaft within the Research Training Group 1765 "SCARE: System Correctness under Adverse Conditions".

References

Berg, J.; Järvisalo, M.; and Malone, B. 2014. Learning optimal bounded treewidth bayesian networks via maximum satisfiability. In *Artificial Intelligence and Statistics*, 86–95.

Cesa-Bianchi, N.; Conconi, A.; and Gentile, C. 2004. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory* 50(9):2050–2057.

Crowley, M.; Boerlage, B.; and Poole, D. 2007. Adding local constraints to Bayesian networks. *Advances in AI* 344–355.

Ellen, C.; Gerwinn, S.; and Fränzle, M. 2014. Statistical model checking for stochastic hybrid systems involving nondeterminism over continuous domains. *International Journal on Software Tools for Technology Transfer*. Published online: 03 August 2014.

Fränzle, M.; Gerwinn, S.; Kröger, P.; Abate, A.; and Katoen, J.-P. 2015. Multi-objective parameter synthesis in probabilistic hybrid systems. In *International Conference on Formal Modeling and Analysis of Timed Systems*, 93–107. Springer.

Fränzle, M.; Gao, Y.; and Gerwinn, S. 2017. Constraint-solving techniques for the analysis of stochastic hybrid systems. In *Provably Correct Systems*. Springer. 9–38.

Fränzle, M.; Hermanns, H.; and Teige, T. 2008. Stochastic satisfiability modulo theory: A novel technique for the analysis of probabilistic hybrid systems. In *International Workshop on Hybrid Systems: Computation and Control*, 172–186. Springer.

Gogate, V., and Dechter, R. 2012. Approximate inference algorithms for hybrid Bayesian networks with discrete constraints. *arXiv:1207.1385*.

Koller, D., and Friedman, N. 2009. *Probabilistic Graphical Models: Principles and Techniques*. MIT press.

MacKenzie, E. J.; Shapiro, S.; and Eastham, J. N. 1985. The abbreviated injury scale and injury severity score: Levels of inter- and intrarater reliability. *Medical care* 823–835.

Murphy, K. P., and Russell, S. 2002. Dynamic bayesian networks: Representation, inference and learning.

Pearl, J. 1985. A constraint propagation approach to probabilistic reasoning. In *Proceedings of the First Conference on Uncertainty in Artificial Intelligence*.

SAE, O., et al. 2014. Taxonomy and definitions for terms telated to on-road motor vehicle automated driving systems. *SAE Standard J3016* 01–16.

Shalev-Shwartz, S.; Shammah, S.; and Shashua, A. 2017. On a formal model of safe and scalable self-driving cars. *arXiv preprint arXiv:1708.06374*.