Towards a Safeguarding Concept for Embedded ML-based Dynamic Models

Hilali Wael
Corporate Research
Robert Bosch GmbH
wael.hilali@de.bosch.com

Neyer Daniel
Corporate Research
Robert Bosch GmbH
daniel.neyer@de.bosch.com

Kloppenburg Ernst
Corporate Research
Robert Bosch GmbH
ernst.kloppenburg@de.bosch.com

Abstract: Machine Learning (ML) methods are gaining popularity in the automotive environment thanks to their capability of complex systems identification throughout different domains. To date, the automotive industry currently relys on the ISO26262 to define the safety measures and development process for conventional physics-based models which are not sufficient for implementation of ML-model for regression tasks. Hence, the aim of this work is to provide a holistic concept of dedicated safeguarding methods, which focuses on ECU implementation aspects of the dynamic ML-model. A systems perspective based on the characteristics of the given model is introduced. The concept requires the development of new customized safeguarding modules for data-based regression tasks. The introduced concept includes off- and online measures, monitoring the input-, output- and the model behavior during runtime, as well as giving a guideline for model training and validation including safety fallback solutions.

1 Motivation and Overview

Neural Networks become an essential tool to approximate highly nonlinear and complex physical relationships for a wide range of applications. To date there is no established development and implementation standard for the safeguarding of black-box recurrent neural networks for regression tasks in safety relevant automotive applications, since such systems have only recently been deployed on real time control units. The automotive industry currently uses the ISO 26262 to define the safety measures and development process for conventional (physics-based) complex models. The recently developed ISO/PAS 21448 (SOTIF) standard specifies a dedicated development process for the analysis, verification and validation of non-faulted scenarios and use cases. However, SOTIF focuses more on L1 and L2 perception tasks, thus autonomous driving tasks, and less on the (recurrent) neural networks for multi-step state regression.
Hence, the aim of this contribution is to provide a concept design of dedicated safeguarding methods for online recurrent Neural Networks, which focus on ECU implementation aspects for the multi-step ahead state prediction for virtual sensors of non-linear dynamic systems. A holistic system perspective based on the known issues and behavior

of the given model characteristics is introduced. This perspective is covered in part by established techniques but also requires the development of new customized safeguarding modules for data-based regression tasks.

1.1 Neural Network Architecture

Since the applications of various Neural Networks (NN) on ECU's are very diverse, this holistic risk and mitigation concept focuses on Neural Network model candidates that are commonly used fo the identification of nonlinear dynamic systems, for example the NN-based regression models that are used as virtual sensors on embedded controller in the automotive industry. Therefore, the presented model is based on a Recurrent Neural Network (RNN) example that incorporates the dynamics of the system into the model structure in order to increase the (dynamic) performance of the prediction.

1.2 Challenges

Based on the given task of implementing the Black Box Neural Network on the ECU, additional implementation aspects including safety features must be considered in order to achieve an Automotive Safety Integrity level (ASIL) according to the intended use and failure risk. In addition to regulations, it is difficult do achieve the necessary degree of safety for a data-based system unless the system itself or a dedicated safeguarding function can control safety relevant operating conditions that result either from the intended use or from inherent model faults. A holistic systems perspective based on the known issues and behavior of the given Neural Network is utilized to safeguard safety throughout the technological development.

The presented safeguarding concept in this work includes offline and online/embedded measures in order to monitor the input-, output- and the model behavior during runtime, as well as to give a guideline for the deep neural network training and validation for critical operating points including physics-based safety fallback solutions. This contribution is not intended to serve as a final statement or minimum or maximum guideline or standard for safeguarding online ML-models. Instead, the intent of this study is to contribute to current activities paving the way for the safe implementation and execution of data-based ML-models on the ECU. The given measures in this report are developed to achieve ASIL-A safety level. In order to guarantee higher safety levels additional features must be implemented according to the intended usage of the model. It is important to note that the concepts elaborated in this work are not coverinng all the safety requirements and not guarantee to reach the safety integrity lebel for the intend use of the embedded function. This contribution is not representing an extensive and complete list of all the measures perfomed by Robert Bosch for the safeguarding of its embedded fucntions and systems.

2 ISO-26262 and SOTIF (Safety of Intended Functionality)

Machine Learning methods and especially Deep Neural Networks become increasingly important based on its capability of abstraction throughout different domains. However, the functional safety standards such as ISO 26262 was not developed with data-based machine learning methods in mind and therefore did not evolve to cover AI integrations. There is currently no established development- and implementation standard for the usage of Black Box Recurrent Neural Networks.

The automotive industry currently uses the ISO 26262 to define the safety measures for conventional (physics based) complex models. The recently developed ISO/PAS 21448 standard (SOTIF) specifies a development process for the analysis, verification and validation of non-faulted scenarios and use-cases and applies to functionality that requires proper situational awareness in order to be safe [WSRA21].

The standard describes the guarantee of safety for the intended functionality in the absence of a fault. This is in contrast with traditional functional safety, which is concerned with mitigating risk due to system failure. Most notably, deep learning algorithms may predict incorrect results. These kinds of limitations are not covered in the ISO 26262 but rather in the recently published ISO 21448 (SOTIF). Therefore, the SOTIF standard provides some valuable risk and mitigation concept approaches that can be applied for regression tasks. However, those approaches do not cover all concerns regarding NN-based regression tasks, which requires additional approaches as shown in figure 1. The appraoches in the following figure are in accordance with the requirements for ASIL-A as an example. Existing standards do not provide solutions to some of the most problematic topics such as stability and extrapolation concerns of a Recurrent Neural Network and self-monitoring issues of Black Box systems, thus safety assurance of AI systems. Therefore, implementing automatic protection and overwatch functions that properly handle critical situation derived from sensor errors, software failures and operational or environmental conditions, must be implemented to move the system to a safe state.

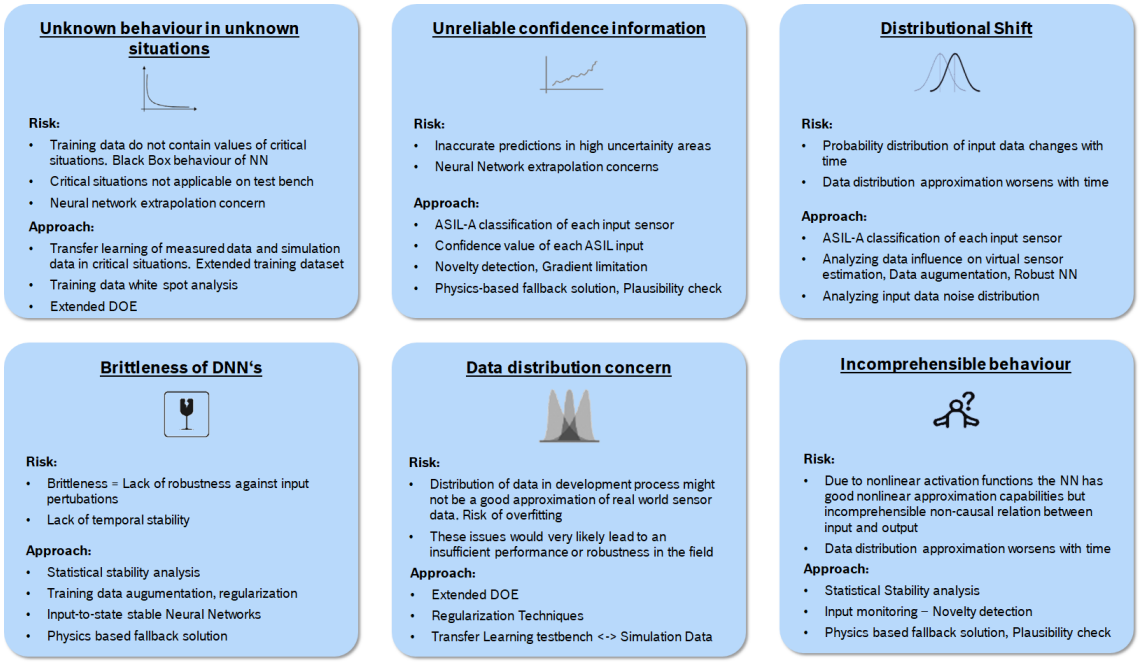

Figure 1: Deep Neural Network risks based on SOTIF evaluation extended to Regression Neural Networks

3 Risks with Recurrent Neural Network Regression Models

Safety related aspects in the automotive area are usually handled through approaches defined in the ISO 26262, the usage of deep learning methods introduces a number of additional safety-related aspects that needs to be covered. Regular Deep Neural Networks do have Black-Box characteristics due to its nonlinear activation functions which enables them to approximate highly non-linear system behavior but also exhibits Black-Box behavior which makes it difficult to evaluate the system inference with respect to safety critical aspects (Issue: Explainability). The Neural Network may face various risks on input-, output- and model level that may lead to inconsistent or unpredictable system behaviour that must be avoided. The safety concerns and potential risks related to Neural Networks for regression tasks are shown in figure 2. The safeguarding methods that need to be implemented, relate to the Neural Network input-, output- and its model structure. The risk- and mitigation concept addressing the individual risks is described in chapter 4.

4 Mitigation Concepts

The recurrent Deep Neural Network can face various risks during the development and deployment on the ECU. In order to guarantee a reliable operation these risks have to be avoid by introducing safeguarding modules which are specialized for the individual Neural Network. The following safeguarding methods relate to a virtual sensor application which includes a recurrent Neural Network design that leads some special stability issues that needs to be considered.

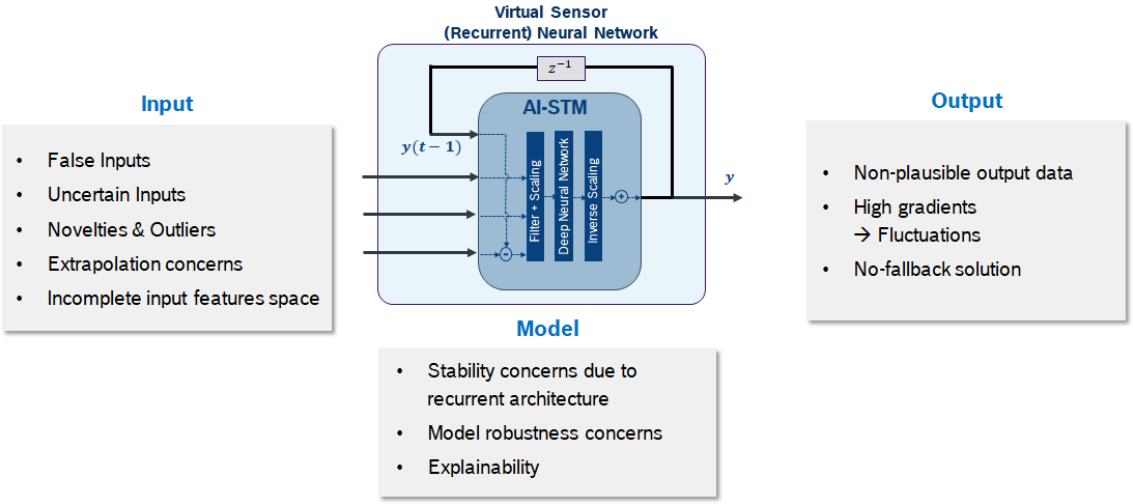

Figure 2: Summary of various risks concerning the safe functionality of the embedded Neural Network categorized into input-, output- and model related risks

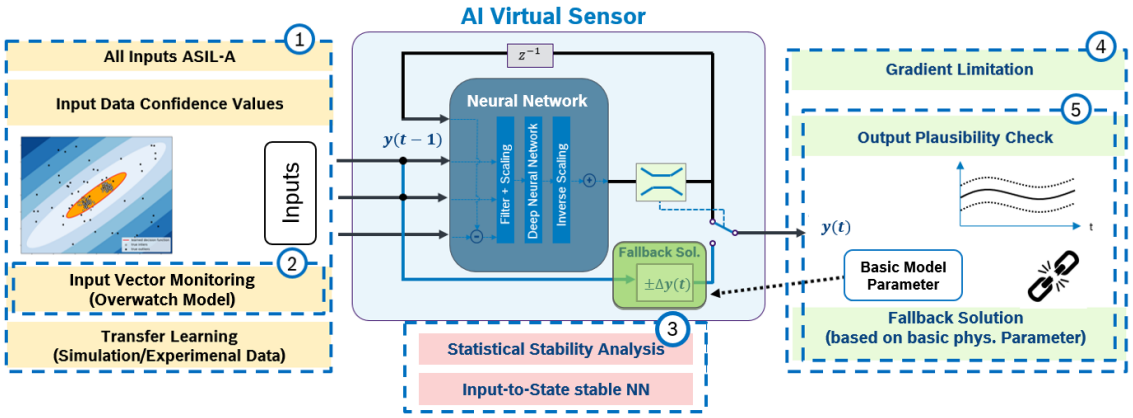

Figure 3: Safeguarding concept including individual modules for Input-, Output- and Model monitoring

4.1 Feature and Signal Quality Requirements

In case the given virtual sensor example has to be deployed on the ECU in a safety critical area of application (fro example assuming the need to ensure the ASIL-A level), the safeguarding concepts must include dedicated safety measures accordingly. In order to be compliant with ISO26262 ASIL-A, all inputs (all individual sensors) must satisfy ASIL-A on a hardware- and software level. According to ASIL classification, all sensor signals must fulfill a certain minimum failure probability and send an additional confidence information for the individual sensor output. Therefore, the safeguarding input monitoring module observes the actual sensor values together with their individual confidence values and compares it with a predefined threshold that triggers a fallback solution in case of sensor failure.

4.2 Input Vector Monitoring - Outlier / Novelty Detection

The performance evaluation is based on the assumption that training and test data are sampled from similar distributions. Deep Neural Networks tend to fail when data distributions during training and operation differ from each other. Therefore, the Neural Network is suceptible to outliers of the nominal distribution given by the training set. The Neural Network can be seen as an universal function approximation with the capability to approximate highly nonlinear system behaviour due to the inherent nonlinear activation function. However, this structural characteristics has the drawback of a black box behaviour which has good interpolation properties, but comes with high uncertainity when the input data is far outside the nominal training data distribution (novelty) as shown in figure 4.

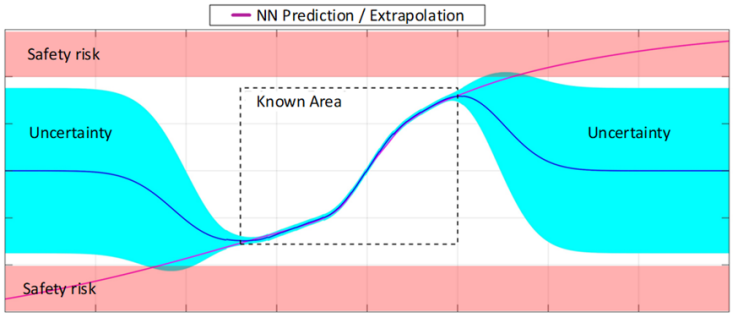

Figure 4: Extrapolation Concern

This is refered to as extrapolation concern. Therefore, a method has to be implemented that monitors the n-dimensional combination of input vectors at each time intervall and evaluates its validity in order to switch to a fallback solution while running on the target control unit (ECU) in realtime.
It is usually not sufficient to define the validity area as the bounding box of the input data. The bounding box is the hyper-rectangle that results from the minima and maxima of the input data per axis direction. The ”data cloud”, i.e., the area of the input space for which training data were available, typically takes a smaller and smaller volume share of the bounding box, the higher the number of dimensions is (e.g., 0.1 % for a current 14d data set). For that purpose, an additional online model will be needed to serve as a one-class classifier, learning the boundary of the input vector hyperspace.
In the case of Gaussian processes, it is known that the validity area can be determined via the model uncertainty (model variance or standard deviation). However, this cannot be calculated or saved on an ECU for memory and computation time reasons. For other ML models, a value for the uncertainty of the model output can usually not be calculated. Several one-class classifiers are available in the literature (see [KM10]) like the Support vector machine, Generalized One-class Discriminative Sub-spaces (GODS), Deep SVDD, AnoGAN and (deep) Autoencoders. These methods are performing differently with different data complexity and higher dimensionality. A key criterion for its application is its suitability to the constrained resources in the ECU, which is strongly dependent on the amount of data, and input vector dimensions.

4.3 Design of Experiment and Transfer Learning

The performance of the Recurrent Neural Network relies on the quality of the training data and is prone to white spots that may not cover critical operating conditions. Although training a Neural Network implies splitting the whole dataset into a training- and test dataset (+ verification dataset) to avoid overfitting and improve generalization of the Neural Network, the closed environment with known and defined operating points allow a certain overfitting for critical operating points that may directly influence safety. That means, the DOE (design of experiment) should include these critical operating areas in order to guarantee a reliable operation on the vehicle. However, these operating areas might be known and well defined but cannot be reached on a testbench. Therefore, a simulation model that is able to simulate all necessary critical operating areas, as well as extending the dataset slightly outside the valid operating areas might benefit the robustness of the Neural Network in case of extreme conditions.For that purpose transfer learning offers a great potential to bridge this gap between simulation and measurement data, and hence filling the ga in the training data distribution.

4.4 Gradient Limitation and Plausibility Check

Possible rapid changes in the predicted virtual sensor output may lead to a abrupt safety critical conditions depending on the system functionality (e.g. a rapid acceleration or deceleration of the vehicle). As a consequence, an output gradient limitation is introduced that inherits the limits of the gradient, based on physical explainable boundary conditions. The gradient limitation function is located before the recurrent feedback in order to modify the NN model behavior towards known operating conditions or settling in case of adverse input distribution or novelties/prediction failure.

4.5 Stability

Although the ISO26262 does not define stability criteria in order to fulfill ASIL-A classification, the stability verification is part of a prior FMEA analysis which identifies where and how the system might fail and evaluates the relative impact of different failures. The presented Explicit Recurrent Neural Network structure improves the prediction performance but has the drawback of stability issues due to the output state feedback. However, for the given Black-Box Recurrent Neural Network, a mathematical global asymptotic stability proof is generally very difficult. For this nonlinear dynamic systems a common energy based autonomous Lypanuov stability approach, which makes use of a Lypunov function $V(x)$ which has an analogy to the potential function of classical systems, cannot be formulated because of the ambiguous modes due to the (appointed) ReLU activation function with n neurons. In general, the output of the activation function can become either mode 0 or a linear mode depending on the neuron output. That leads to a total number of 2^n different ambiguous modes. In order to analyze the stability of the Neural Network, all individual modes have to be analyzed which is not feasible for the given design. The

second problem is, that just because all individual modes are stable does not mean that the combination of all modes are stable and vice versa. Therefore, a mathematical global asymptotic stability is very difficult for the current Explicit Recurrent Neural Network design, at least a posteriori. In the last few years, several methods to build by-design stable recurrent models and proove its stability mathametically a priori are reported in the literature [BTFS20].

An alternative approach for this special type of Neural Network describes a data-based approach which attempts to provide a statement about the statistical probability of the instability by using high amount testbench-, simulation and field-test data. This approach requires data that that covers all operating points including critical conditions. This leads to an increased effort for a suitable design of experiment. In order to analyze the statistical probability of instability an oscillation detection algorithm must be integrated in order to identify limit cycles or unstable behaviour.

In order to overcome concerns regarding stability and explainablity, modern hybrid Neural Network solutions can be applied which are based on prior physical knowledge that can be incorporated into the system design and training process. This solution features an approach that enables a stable-by design process which includes a proof of stability based on input-to-state stable hybrid model design.

4.6 Fallback Solution

The detection of an unusual input-, output- or system state in combination with confidence values allows for a number of established safety procedures to be used for an embedded Neural Network such as giving more weight to parallel information path or switching to a failsafe/fallback solution in case of failure. This fallback solution includes a basic physical representation of the given system by introducing a reduced simulation model or a low dimensional LUT (look-up-table) that acts as a backup in case of Neural Network failure or force the system into a predefined state. The Safeguarding Solution does not provide a highly accurate representation, but provides physically motivated upper and lower bounds of the virtual sensor output gradient, therefore also acting as a plausibility check for the Neural Network prediction.

5 Summary

In this contribution, we have presented a concise list of different safety concerns regarding the deployment of recurrent neural network as an online AI function in the ECU for the identification of non-linear dynamical systems. We have discussed some of the possible mitigation approaches addressing those safety concerns, with a special focus on its suitability for embedded use cases. It is important to note at this stage that the discussed approaches have very different maturity and complexity levels. While all of the approaches can definitely contribute to a safety case, for the time being it remains an open question when a specific safety concern is sufficiently mitigated. This contribution is trying to participate to the efforts in the automotive community to pave the way for the

safe deployment of online AI functions. Thus, it is essential to collect more knowledge and consolidate it in standardization activities in order to define suitable processes and best-practices.

References

[BTFS20] Fabio Bonassi, Enrico Terzi, Marcello Farina, and Riccardo Scattolini. LSTM Neural Networks: Input to State Stability and Probabilistic Safety Verification. In *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, volume 120 of *Proceedings of Machine Learning Research*, pages 85–94. PMLR, 10–11 Jun 2020.

[KM10] Shehroz S. Khan and Michael G. Madden. A Survey of Recent Trends in One Class Classification. In Lorcan Coyle and Jill Freyne, editors, *Artificial Intelligence and Cognitive Science*, pages 188–197, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[WSRA21] O. Willers, S. Sudholt, S. Raafatnia, and S. Abrecht. Safety Concerns and Mitigation Approaches Regarding the Use of Deep Learning in Safety-Critical Perception Tasks. *ArXiv e-prints*, January 2021.