Managing Uncertainty of AI-based Perception for Autonomous Systems

Maximilian Henne, **Adrian Schwaiger** and **Gereon Weiss**
Fraunhofer ESK, Hansastr. 32, 80686 Munich, Germany
{firstname.lastname}@esk.fraunhofer.de

Abstract

With the advent of autonomous systems, machine perception is a decisive safety-critical part to make such systems become reality. However, presently used AI-based perception does not meet the required reliability for usage in real-world systems beyond prototypes, as for autonomous cars. In this work, we describe the challenge of reliable perception for autonomous systems. Furthermore, we identify methods and approaches to quantify the uncertainty of AI-based perception. Along with dynamic management of the safety, we show a path to how uncertainty information can be utilized for the perception, so that it will meet the high dependability demands of life-critical autonomous systems.

1 Introduction

The use cases for autonomous systems have long been limited to highly controlled environments where mostly deterministic algorithmic approaches were applied for perception. This changed after *deep neural networks (DNNs)* proved to be very advantageous in classification and regression tasks, beating previously used shallow methods for recognizing patterns in various fields like computer vision, speech recognition, and many more. As a consequence, DNNs are now given large responsibility in perceiving and interpreting the environment for highly autonomous systems like self-driving vehicles. Nevertheless, especially adversarial attacks have shown that present confidence scores of DNNs do not correspond to reliable uncertainty estimates of a prediction. In order to ensure sufficient safety for these systems, the uncertainty of DNNs needs to be quantified for giving reliable confidence values over the outputs as system planning and control functionalities of autonomous systems strongly rely on this information. For instance, if a computer vision based DNN responsible for detecting pedestrians is indicating high uncertainty in predicting the presence or absence of an immediate obstruction, the vehicle may follow alternative strategies to reduce the uncertainty or take measures that the uncertainty is not endangering its safety, e.g. by slowing down or planning other trajectories.

Unlike formal safety verification approaches, which provide safety guarantees for certain input regions, the approach we envision encompasses developing reliable methods to capture uncertainties in deep learning through combining methods which tackle DNN confidence and to evaluate them in simulated and real-world driving scenarios for autonomous vehicles.

2 Related Work

Recent advances towards autonomous vehicles make heavy use of cameras, radars, and lidars, whose outputs are then interpreted by a machine learning model to perceive their surroundings, e.g. detect other traffic participants, lane markings, etc. Popular DNNs for these tasks include ResNet and DenseNet for image recognition, Faster R-CNN and YOLOv3 for real-time object detection, and PointNet++ and VoxelNet for object detection on point clouds. However, although these networks achieve impressive results on their respective test datasets, the question arises whether they are sufficiently reliable for safety-critical decisions, like autonomous vehicles [Koopman and Wagner, 2017; Frtunikj and Fürst, 2019]. Adversarial attacks show exemplary that DNNs are vulnerable and overconfident when dealing with out-of-distribution examples [Goodfellow *et al.*, 2014; Kurakin *et al.*, 2017; Liang *et al.*, 2017]. Efforts to take measures against these weaknesses are subject to current research [Madry *et al.*, 2017; Papernot *et al.*, 2016], but are not sufficiently comprehensive to enable using DNNs in safety-critical applications. Therefore, it is important to at least estimate the uncertainty of a DNN. Some of the most promising approaches towards dealing with and measuring uncertainty in DNNs are discussed in detail in Section 4.1

3 Perception of Autonomous Systems

For an autonomous system's operation in different scenarios, many decisions rely on its perception of the environment. An example are autonomous vehicles which must drive safely in changing surroundings and complex situations. As AI is capable to abstract from distinct learned scenarios to solve present tasks, DNNs are commonly used in different stages of sensing and interpreting a system's environment. There is the possibility to implement all tasks by DNNs *end-to-end* or solving single stages of the perception in a *modular* approach. The latter (cf. top of Figure 1) allows utilizing deterministic and verifiable algorithms for single tasks (e.g. depth detection or actuating) and is currently favored from a safety perspective.

However, a major open challenge for ensuring safe behavior of an autonomous system is to keep its failure rate below unaccepted risk levels. Just as an exemplary indication, for automotive systems' functional safety with highest safety integrity level this is 10^{-9} failures/hour. Unfortunately, today's utilized DNNs do not allow for reliable classification which meets the high safety requirements needed for autonomous systems. A major factor is the uncertainty of a DNNs' output. However, a reliable perception is an inevitable precondition for all subsequent decisions. Traditionally, a common practice to integrate such unreliable components into dependable systems is to encapsulate and monitor them by a so-called *safety envelope*. In case of a detected fault, the untrusted component is isolated and a verified safety path is used. For perception such an approach is very challenging, since there is currently no alike powerful deterministic and safe substitute available. Thus, the safe backup would be to shutdown or transit to a less automated behavior, which results in a system which is safe but with low availability. Therefore, we need high resilience of the system, i.e. ensuring dependability while optimizing performance.

To achieve this, a solution for limiting the uncertainty of AI-based perception under bounds which are acceptable for safety-relevant applications is required. Any information on the uncertainty of the individual perception tasks may help to improve the supervision of an inherent unsafe AI. For instance, derived high uncertainty of a region may trigger a cross-validation or recheck with additional measures, like other sensors or deterministic algorithms. Hence, in the following we introduce suitable approaches to quantify the uncertainty of DNNs, which may help to pave the way towards safe perception.

4 Taming Uncertainty of AI-based Perception

4.1 Uncertainty in Deep Neural Networks

For the context of this paper, uncertainty is defined as a state of limited knowledge about the correctness of a predicted outcome. In general, we can distinguished two kinds of uncertainty in DNNs (cf. [Kendall and Gal, 2017]). *Epistemic uncertainty*, which is also referred to as model uncertainty, accounts for uncertainty in model parameters and captures the ignorance of which model generated the collected data. This kind of uncertainty would be zero if the model was trained with all possible data existing which is obviously unfeasible. Epistemic uncertainty is modeled by placing a prior distribution of the weights of a model and capturing how these weights vary given some data. *Aleatoric uncertainty*, which is also referred to as data uncertainty captures the noise inherent in the observations. This kind of uncertainty is modeled by placing a distribution over the outputs of a model.

Calibration

One approach is to calibrate the outputs of a DNN's softmax-layer so that their prediction weights match the probabilities of being the correct classes respectively, meaning that a class predicted with weight p is correct p percent of the time. Large increases in model capacity and complexity of DNNs during the last years like rising depth, the use of more convolutional filters, or the use of batch normalization have been identified to negatively affect model calibration [Guo *et al.*, 2017] which often leads to overconfident predictions. For calibrating these DNNs, [Guo *et al.*, 2017] suggest temperature scaling as straightforward method for minimizing the *expected calibration error*. Even though this approach significantly improves model calibration it still does not present a complete solution to the uncertainty problem, as the networks are calibrated relative to a dataset. Outside of this distribution the network is not calibrated. As a result, calibration based methods alone are not suitable for quantifying uncertainty in DNNs for perception, as the data for testing often varies very much from the training data. Nevertheless, calibration can be beneficial, if very large datasets are used for training, which are more representative for the overall data generating distribution. This can be the case in autonomous systems that operate in more controlled environments, like logistic or production facilities. Furthermore, the calibration of a DNNs' predictions provide a well suited measure to evaluate other approaches dealing with reliable uncertainty estimation.

Out-of-Distribution Detection

Another approach to tackle uncertainty is to detect *out-of-distribution (OOD)* inputs by identifying data which is very different from the training data. One way to select them is by utilizing temperature scaling ([Guo *et al.*, 2017]) in combination with adding controlled perturbations (cf. [Goodfellow *et al.*, 2014]) to the input. This is done by using a pre-trained classifier, in order to enlarge the softmax score gap between in- and out-of-distribution examples and apply simple thresholding ([Liang *et al.*, 2017]). However, the performance of this method is highly depending on this pre-trained classifiers. It can be problematic, if the classifier does not separate the maximum value of the predictive distribution well enough for in- and out-of-distribution inputs. This problem can be reduced by changing the loss function to additionally minimize the *Kullback-Leibler (KL)* divergence from the predictive distribution on out-of-distribution samples to the uniform distribution, in order to give less confident predictions on these points ([Lee *et al.*, 2017]). Nonetheless, low confident posteriors over classes could indicate uncertainty in the prediction as a result of a high overlap region of class in an in-distribution input (*data uncertainty*) or an out-of-distribution input far from the training data (*distributional uncertainty*). In order to deal with this problem, [Malinin and Gales, 2018] present a novel framework of so called *Prior Networks* which parameterize a prior distribution over predictive distributions in order to allow data, distributional and model uncertainty to be comprehended within a consistent interpretable framework.

OOD networks can be useful for detecting adversarial attacks or recognizing classes that are not contained in the model, which is highly relevant regarding many safety-critical perception tasks. Nonetheless, OOD methods often do not provide reliable scores of the uncertainty of the prediction itself. Rather they predict the probability of a data point being an OOD input or not. However, the approach using Prior Networks [Malinin and Gales, 2018] is able to extract different kind of uncertainties in one model, which makes this method interesting for application in perception.

Bayesian Neural Networks

Probably the most popular family of approaches to capture uncertainty in deep learning are *Bayesian Neural Networks (BNN)*, which integrate uncertainties in form of probability distributions over its weights. Unfortunately, it is intractable to evaluate the posterior probability of a DNN analytically due to the fact that it requires integration over all possible model parameters. To deal with this problem, two kinds of approaches are used: *variational inference* and *sampling methods*. Variational inference methods try to approximate this true posterior over the model parameters with a different distribution from a tractable family (e.g. gaussian) by finding the parameters of this distribution that minimize the KL divergence to the true distribution. The problem with variational inference is that it has a very high bias, as we manually choose the distribution family which the weights are drawn from. On the other hand, sampling methods approximate the true distribution with the average of samples drawn from it. One way to do it is by using a *Markov Chain Monte Carlo (MCMC)* algorithm, which constructs a markov chain with the desired distribution as its equilibrium distribution. However, even though in theory MCMC would lead to a perfect approximation of the true posterior, it requires way too much computation for most DNNs to converge within acceptable time.

A more recent theoretical finding provides a Bayesian interpretation of the regularization technique known as *dropout* [Gal and Ghahramani, 2016]. The argument is that dropout could be used for performing variational approximation of a BNN with a Bernoulli distribution prior from which *Monte Carlo (MC)* sampling is done. In practice, this finding provides an easy way to turn a conventional DNN into a BNN by simply applying dropout during training and testing time. It averages over multiple forward passes through the DNN during testing time, in order to achieve a MC approximation of the predictive distribution. Due to its simplicity, high scalability, and good generalization performance, this approach is widely used to tackle the problem of deriving reliable uncertainty estimation of DNNs. For instance, [Kendall *et al.*, 2017] used this method for capturing pixel-wise uncertainty on the perception task of semantic segmentation. In order to capture epistemic and aleatoric uncertainty separately in a single model for pixel-wise semantic classification, [Kendall and Gal, 2017] successfully use MC dropout for model uncertainty together with placing a gaussian distribution over the logits (i.e. values for each class label before applying the softmax function) with their respective values as mean and additional noise as variance for data uncertainty.

Overall, many state-of-the-art BNNs utilize MC dropout to capture uncertainty. The main reasons are that it can be implemented easily in any given DNN without many changes and the strong results achieved by this method. Furthermore, [Gal *et al.*, 2017] propose a new dropout variant which gives improved calibration and performance. Nevertheless, it requires many forward passes on a single input (often more than 50) to obtain a principled estimate of the predictive distribution. This is a problem for real-time perception tasks, like vision based object detection or Lidar-data based segmentation, for which computation resources and time are limited. But despite the slow run-time, MC dropout presents a well suited method for dealing with uncertainty in perception tasks that are not sensitive to performance limitations.

Deep Ensembles

Another non-Bayesian approach to quantify uncertainty uses ensembles of DNNs for obtaining a distribution over the predictions. [Lakshminarayanan *et al.*, 2017] adopt an ensemble of DNNs together with adversarial training [Goodfellow *et al.*, 2014] to smooth the predictive distribution. The ensembles are used as uniformly-weighted mixture to get the predictive mean and variance associated with the prediction. This method is actually similar to MC dropout in the sense that both methods sample from many different models. The key difference is that MC dropout only samples from sub-models of the initial model which share weights. As a consequence, *Deep Ensembles (DEs)* also capture model uncertainty by averaging predictions over multiple models and furthermore, due to the use of adversarial training, allows for increased robustness to model misspecification and out-of-distribution examples. As sampling is also involved in DEs, they share the same weaknesses regarding run-time performance with MC dropout. Nevertheless, DEs deliver highly competitive results, which have been shown to be at least on par with MC dropout. Due to simplicity and strong performance,DEs provide an auspicious approach to tackle uncertainty in perception tasks.

4.2 Dynamic Management of Perception Uncertainty

The diverse advances for estimating uncertainty of DNNs described in the previous Section 4.1 seem very promising. Nevertheless, up to now there is no solution which is capable to enable adequate perception reliability. Therefore, we believe that only a combination approach will be successful, which uses different methods for deriving uncertainty information in the various tasks of the functional perception chain (see Figure 1). In addition to a safety envelope architecture [Weiss *et al.*, 2018], this will provide improved reliability of the DNNs' performance. However, as it is not likely that

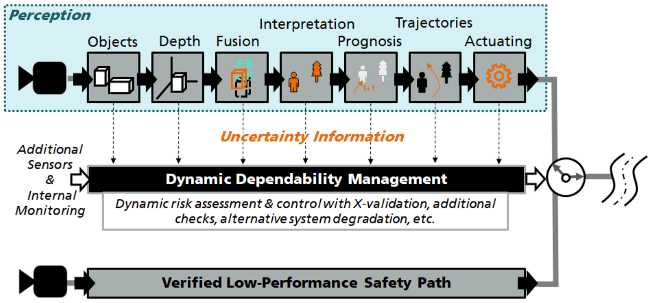
Perception
Objects
Depth
Fusion
Interpretation
Prognosis
Trajectories
Actuating
t+1
Uncertainty Information
Additional Sensors & Internal Monitoring
Dynamic Dependability Management
Dynamic risk assessment & control with X-validation, additional checks, alternative system degradation, etc.
Verified Low-Performance Safety Path

Figure 1: Concept overview for utilizing uncertainty information of modular perception stages for dynamic dependability management.

this alone will meet the high requirements of safety-critical systems with respect to fault rates, we propose to assess the actual risks dynamically considering the uncertainty information and to take respective safety measures at runtime [Trapp *et al.*, 2018]. As example we consider the case that the uncertainty of a detected obstacle is low and the expected severity of a wrong classification is high, e.g. as an impact would be likely. The dynamic dependability management would initialize a cross-validation of that area with an additional sensor and algorithm, which is not subject to the same type of weaknesses. Additionally, by dynamically assessing and managing the risks, environment regions can be classified and treated differently with respect to their characteristics and impact on the systems's safety. For instance, a higher uncertainty of the perception in a specific area could be acceptable, if this region is not likely to have a substantial impact on the system's safety, e.g. it is not in the trajectory path of the autonomous system and it is unlikely that an object will be able to move from this area into the driveway.

Overall, the combination of modularly estimating the uncertainty of perception and dynamic dependability management poses a promising solution for overcoming the challenge of unreliable perception. In the scope of our work, we will examine several presented methods for estimating uncertainty of different modules involved in the perception chain and identify the most suitable for each case.

5 Conclusion and Future Work

Reliable perception for autonomous systems poses an open challenge. With a comprehensive overview, we showed that various methods capturing uncertainty in DNNs may be utilized to improve confidence in AI-based perception. Overall, we identified Prior Networks, Concrete Monte Carlo dropout, and Deep Ensembles as the most promising approaches for deriving reliable confidence scores. By exploiting such uncertainty information in the different stages of an autonomous systems' perception, a dynamic management with the given uncertainty becomes feasible.

In our ongoing research we develop an approach for reliable perception, based on the combination of systematically deriving perception weaknesses, architectural designs, and dynamic dependability management.

Acknowledgments

This work has been partially funded by the Bavarian Ministry of Economic Affairs, Regional Development and Energy.

References

[Frtunikj and Fürst, 2019] J. Frtunikj and S. Fürst. Autonomous vehicle safety: An interdisciplinary challenge. *Twenty-seventh Safety-Critical Systems Symposium, Bristol, UK*, 2019.

[Gal and Ghahramani, 2016] Y. Gal and Z. Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *33rd International Conference on Machine Learning*. PMLR, 2016.

[Gal *et al.*, 2017] Y. Gal, J. Hron, and A. Kendall. Concrete dropout. In *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc, 2017.

[Goodfellow *et al.*, 2014] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. 2014.

[Guo *et al.*, 2017] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *34th International Conference on Machine Learning*, 2017.

[Kendall and Gal, 2017] A. Kendall and Y. Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems 30*. 2017.

[Kendall *et al.*, 2017] A. Kendall, V. Badrinarayanan, and R. Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. In *British Machine Vision Conference*, 2017.

[Koopman and Wagner, 2017] P. Koopman and M. Wagner. Autonomous vehicle safety: An interdisciplinary challenge. *IEEE Intelligent Transportation Systems Magazine*, 9(1):90–96, Spring 2017.

[Kurakin *et al.*, 2017] A. Kurakin, I. Goodfellow, and S. Bengio. Adversarial examples in the physical world. In *Workshop Track of the International Conference on Learning Representations*, 2017.

[Lakshminarayanan *et al.*, 2017] B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems 30*. 2017.

[Lee *et al.*, 2017] K. Lee, H. Lee, K. Lee, and J. Shin. Training confidence-calibrated classifiers for detecting out-of-distribution samples. 2017.

[Liang *et al.*, 2017] S. Liang, Y. Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks, 2017.

[Madry *et al.*, 2017] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks, 2017.

[Malinin and Gales, 2018] A. Malinin and M. Gales. Predictive uncertainty estimation via prior networks. In *Advances in Neural Information Processing Systems 31*. Curran Associates, Inc, 2018.

[Papernot *et al.*, 2016] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE Symposium on Security and Privacy*, 2016.

[Trapp *et al.*, 2018] M. Trapp, G. Weiss, and D. Schneider. Towards safety-awareness and dynamic safety management. In *14th European Dependable Computing Conference*, 2018.

[Weiss *et al.*, 2018] G. Weiss, P. Schleiss, D. Schneider, and M. Trapp. Towards integrating undependable self-adaptive systems in safety-critical environments. In *13th International Symposium on Software Engineering for Adaptive and Self-Managing Systems*, 2018.