Real-Time Explainable Plausibility Verification for DNN-based Automotive Perception*

Mert Keser[1,2]

[1]Technical University of Munich, Germany
[2]Continental AG, Germany

Abstract
Deep Neural Network (DNN) based perception algorithms have demonstrated remarkable success in tackling perception problems. However, their inherent shortcomings, such as brittleness, opacity, and unpredictable responses to edge cases, complicate the verification of their decision-making processes. This challenge stems from the DNN-based algorithms' lack of transparency and interpretability, making it difficult to assess their plausibility against existing environmental knowledge. For safety-critical applications like autonomous driving (AD), monitoring the decisions of DNN-based perception algorithms in relation to established environmental knowledge during operation is crucial. This paper reviews current methods for verifying the plausibility of DNN-based perception algorithms and proposes innovative approaches for real-time, explainable plausibility verification within the context of DNN-based automotive perception.

Keywords
Explainable AI, real-time plausibility, autonomous vehicles

1. Introduction

Recent advancements in Deep Neural Networks (DNNs) have broadened the scope of DNN-centric perception methods in safety-critical applications such as autonomous vehicles [1]. The real-world deployment of these DNNs, however, raises significant concerns regarding potential safety violations, which may impede the approval and conformance of vehicles to safety standards [2, 3]. This situation also prompts questions about the safety of such approaches from a societal perspective. DNNs, when deployed, are expected to effectively handle complex, corner-case scenarios that generate data not encountered in the training set. Under such conditions, there is often a substantial drop in DNN performance [4, 5]. Given this context, the plausibility of a DNN is defined as the degree to which its outputs or decisions align with established domain knowledge or expected behavior. As comprehensive performance evaluation of DNNs cannot be attained by exhaustively listing and assessing every conceivable scenario the DNNs might encounter, implementing plausibility verification methods becomes crucial to guarantee the accuracy and safety of perception functions in AD.

Late-breaking work, Demos and Doctoral Consortium, colocated with The 1st World Conference on eXplainable Artificial Intelligence: July 26–28, 2023, Lisbon, Portugal

*Corresponding author.

mert.keser@tum.de (M. Keser)
0000-0001-7373-437X (M. Keser)

The primary reason for the performance drop in DNNs during real-world deployments stems from the methodology employed in training and testing these networks. Typically, DNNs are developed using large datasets, which are partitioned into training and testing subsets. The common practice is to train these networks using the training subset and validate them with the testing subset of the datasets [6]. It is generally assumed that the test data accurately reflects the DNNs' performance during a real-world deployment. However, as the training and testing datasets are derived from the same distribution, evaluating DNNs with test data may not truly reflect their performance under real-world conditions. In the real-world deployments, DNNs encounter edge-case scenarios [7] and may even face adversarial attacks [8], making it difficult to predict their performance based solely on the testing subset of the dataset. In the case of AD, uncertainties arise from environmental influences, sensing hardware, and perception software [9]. These uncertainties consequently generate ambiguity in the classification output of DNNs. Therefore, there is of utmost importance to continuously monitor the plausibility of the decisions made by DNNs in the real world. Additionally, for the real-world deployment of DNNs in AD, system-level measures such as error identification and treatment are also required, as reflected in the upcoming AD safety standards ISO/TR 4804 [2].

In the context of autonomous driving, instant feedback on the decision-making process of Deep Neural Networks (DNNs) is also crucial for building user trust and ensuring safety. Providing users with immediate explanations for the vehicle's actions, such as sudden braking or lane changes, enhances their understanding and confidence in the technology. This feedback becomes particularly essential in situations where the DNN detects objects or events that may not be immediately apparent to human observers. By offering transparent and interpretable explanations for these decisions, users can gain insights into the DNN's behavior and develop a better sense of trust and comfort in relying on autonomous systems. This instant feedback enables users to have a deeper understanding of why specific decisions are made, fostering acceptance and facilitating effective collaboration between humans and autonomous systems.

One significant challenge in verifying the plausibility of decisions made by Deep Neural Network (DNN) models lies in their inherent limitations. Often referred to as 'black-boxes', these models are characterized by a critical lack of transparency and interpretability [10]. The lack of transparency and interpretability prevents the validation of perception functions in AD. Consequently, there has been a surge of research interest in the explainability of DNN models employed in AD [9]. However, to the best of the author's knowledge, the existing explainability methods have limitations, including an inability to operate in real-time. As AD involves high-stakes decisions, it is of paramount importance to evaluate the decisions made by DNNs during their operation. Therefore, a plausibility verification method must function in real-time and verify the plausibility of DNNs' decisions in AD systems.

In this Ph.D. research, my objective is to address the need for real-time plausibility verification methods within the context of DNN-based automotive perception. The methods developed through this research will be able to validate decisions made by DNNs, in relation to the knowledge available about the surrounding environment.

2. Research Problem

Autonomous driving has emerged as one of the most revolutionary advancements in transportation, promising improved safety, convenience, and efficiency [11, 12]. However, it demands a high degree of robustness and reliability, especially in the operation of Deep Neural Networks (DNNs) in perception functions. To effectively manage critical situations that may arise during deployment, these DNNs must demonstrate exceptional proficiency and accuracy in perceiving and interpreting their environment. Moreover, the safety of autonomous vehicles can be further ensured by verifying the decisions made by these networks. In the light of these challenges, this Ph.D. research aims to investigate methods for real-time plausibility verification of DNN-based perception functions in autonomous driving and propose strategies to enhance their safety capabilities. Specifically, the overall aim of this project is to answer the following question:

- How can the **plausibility** and the **safeguarding** of the **DNN-based perception functions** be improved?

Since the area of real-time plausibility verification for DNNs is still largely unexplored, the following aspects have been chosen to define the scope of the research. These aspects include:

- Plausibilizing decisions of DNNs in high-stake applications
- Automating plausibility checks for DNNs in perception functions for autonomous driving
- Implementing real-time plausibility checks for DNNs

Based on these aspects, we propose the following research questions to investigate the previously mentioned areas thoroughly:

RQ1 How can we evaluate the outputs of DNN-based perception functions for their conformity with existing knowledge?

RQ2 How can we implement real-time plausibility verification methods within the context of DNN-based perception functions?

RQ3 How can we automatically validate the plausibility of a DNN's decision?

This Ph.D. research, which began in May 2022, focuses on investigating real-time plausibility checks for DNNs within the perception function of autonomous vehicles. Section 3 identifies key related works that define the research problem, while Section 4 discusses the methods proposed to tackle the issues outlined in the research questions. Lastly, Section 5 concludes the paper and provides an overview of the planned thesis work.

3. Related Work

3.1. Explainable AI in Autonomous Driving

Explainability in AD is essential for safety and user trust. Deep neural networks (DNNs) used in AD perception functions often lack transparency, making explainability increasingly crucial for identifying corner cases and potential failure modes, as well as building user trust in the system [13, 14]. Explainable artificial intelligence (XAI) methods for AD can be categorized into local explanations, global explanations, and explainable driving models.

Local explanations can be categorized into saliency-based and counterfactual explanations. Saliency-based methods identify important image regions that contribute to the output of DNNs, using techniques like back-propagation [15, 16], perturbation [17], or local-approximation [18]. Counterfactual explanations find minimal input modifications to change the model's decision, helping to understand model sensitivity and robustness [17, 19].

Global explanations aim to provide a comprehensive summary of a model's behavior. Model translation approaches aim to make opaque models interpretable by converting them into more transparent models [20, 21]. Representation explanation techniques offer insights into the information captured within the model's internal data structures at varying levels of granularity, improving our understanding of model behavior [13, 22].

An emerging trend is incorporating explainability directly into the model architecture, designing intrinsically explainable models. Drawing inspiration from modular systems, examples in this domain include the end-to-end architecture of ChauffeurNet [23].

In conclusion, explainability in AD is vital due to its safety-critical nature and increased use of DNNs. While local methods offer valuable insights into specific cases, they have limitations in terms of verifiable constraints. Global methods provide comprehensive summaries of model behavior but can be challenging to implement in real-time. As a result, new XAI techniques are needed for real-time plausibility verification in autonomous driving, ensuring safety and user trust in this fast-evolving field.

3.2. Plausibility Verification of DNNs

Many studies have aimed at assessing the validity of decisions made by DNNs. One common approach involves using uncertainty estimation to verify the output of the decision by DNNs. For instance, Feng et al. [24] utilize uncertainty estimates of DNNs to compute a trustworthiness score. Gaussian YOLOv3 [25], on the other hand, identifies false positives by calculating bounding box localization errors. Autoencoders have also been employed to examine false positives in DNNs [26]. In a recent study, the output of 2D object detectors has been verified with human-interpretable concepts by using concept-based interpretations [27].

A prevalent alternative method involves validating the decisions made by DNNs using other sensor modalities. For example, Khesbak et al. [28] and Geissler et al. [29] utilize sequential checking methods to evaluate the consistency of object detection across multiple sensor streams. Additionally, Vivekanandan et al. [3] employ energy-based optimization techniques to verify object detection across various sensor streams.

In conclusion, various studies have sought to verify the plausibility of DNN decisions using approaches like uncertainty estimation and alternative sensor modalities. However, these approaches must address the critical need for explainability in DNNs. As such, there is an increasing demand for methods that validate DNN decisions and enhance interpretability, fostering deeper insight into their decision-making. Therefore, bridging the gap between plausibility verification and explainability in DNNs is essential.

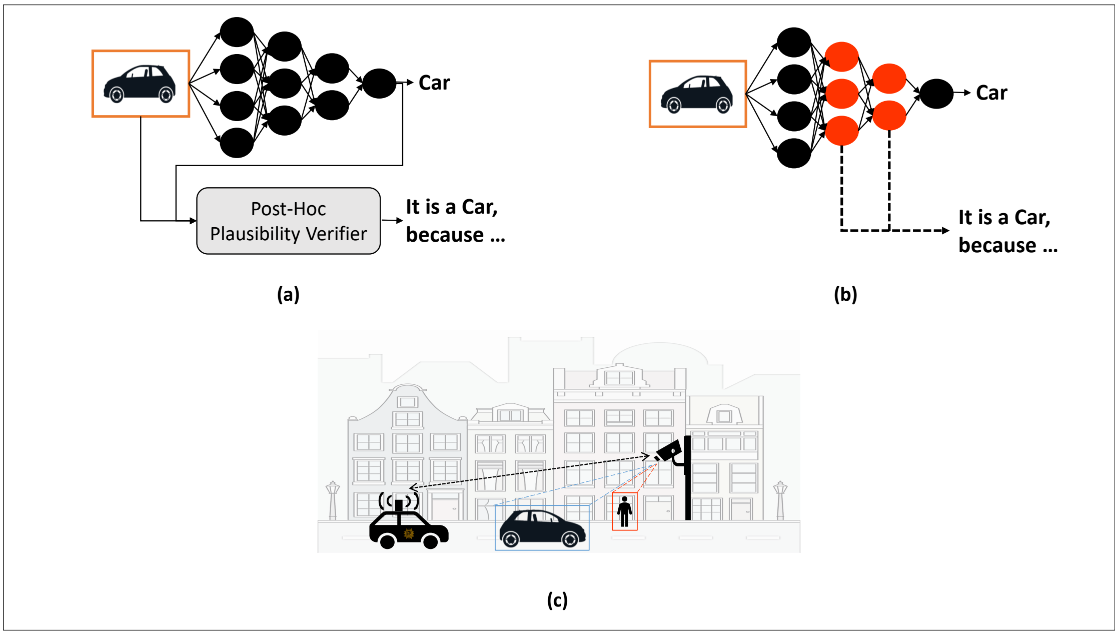

Figure 1: Overview of the three-stage Ph.D. research structure. (a) Post-hoc real-time plausibility verification method, (b) Ante-hoc plausibility verification method, and (c) Automatic explainable plausibility verification using collaborative perception.

4. Proposed Solutions

To address the identified research questions in Section 2, three distinct studies are presented. To the best of our knowledge, this is the first work that concentrates on real-time plausibility verification with explainability features. The Ph.D. research has been structured into three stages to address these research questions:

1. Development of a post-hoc, real-time plausibility checker
2. Development of ante-hoc XAI methods for a plausibility check
3. Explainable plausibility verification of perception DNNs using collaborative perception

The initial stage of the Ph.D. research involves conducting a literature review on XAI methods, assessing their real-time performance and suitability for verifying the plausibility of DNN's behavior. The selected method will function as a parallel plausibility verifier for the network's decisions. This first research stage will address RQ1 by proposing methods that examine DNN behavior in accordance with existing knowledge and RQ2 by investigating real-time plausibility verification.

The subsequent stage of the Ph.D. research focuses on integrating the chosen XAI method into the DNNs' network architecture. While the initial step is suitable for understanding false positive detection in DNNs, it falls short in identifying false negative detections. Since the parallel plausibility verifier is only activated by the DNN's output, it can solely detect false positive cases. By incorporating the plausibility verifier into the DNNs' network architecture, the research aims to minimize false negative and false positive detections.

The final stage of the Ph.D. research centers on the automation of plausibility verification. This stage addresses RQ3 by proposing methods to automate real-time, explainable plausibility verification. strategy is utilizing a collaborative perception environment to verify the plausibility of perception functions in autonomous vehicles. In collaborative perception, multiple views of the environment offer an enhanced understanding of the surroundings. This improved comprehension of the environment can be harnessed to validate the behavior of DNNs.

Concept bottleneck models [30] present a promising avenue for real-time plausibility verification in DNN models. These models essentially offer a simplified depiction of the knowledge captured by the DNN, acting as intermediaries that transform the complex internal representations of DNNs into more comprehensible outputs for users, which aids explainability.

However, bottleneck models are limited by their reliance on the presence and quality of interpretable features [31]. This reliance can curb their application in scenarios where these features are sparse or absent. Additionally, their oversimplification of the DNN's knowledge can sometimes result in the loss of important details.

Despite these challenges, bottleneck models hold considerable potential for enhancing real-time plausibility verification. Their capability to transform intricate DNN representations into interpretable ones could play a crucial role in developing real-time, explainable plausibility checks. Therefore, the combination of bottleneck models and plausibility verification forms a compelling area of exploration for this research.

5. Conclusion

This paper has identified the need for real-time, explainable plausibility verification methods for DNN-based automotive perception systems. We have proposed a three-stage approach to address this need. Our approach includes the development of a post-hoc real-time plausibility checker, an ante-hoc XAI method for plausibility checking, and an explainable plausibility verification process using collaborative perception.

The Ph.D. research presented in this paper will contribute to the field by developing novel methodologies for real-time plausibility verification with explainability features. As a result, the proposed methods have the potential to significantly enhance the safety and reliability of autonomous vehicles and advance the deployment of DNN-based perception algorithms in real-world applications. Ultimately, this research aims to bridge the gap between plausibility verification and explainability in DNNs, paving the way for more robust and trustworthy AI systems in safety-critical domains.

Acknowledgement

The research leading to these results is partly funded by the German Federal Ministry for Economic Affairs and Climate Action within the "KI Wissen" project. Special thanks are given to Gesina Schwalbe and Alois Knoll for their invaluable guidance during the first year of the Ph.D. program.

References

[1] Y. Huang, Y. Chen, Autonomous driving with deep learning: A survey of state-of-art technologies, arXiv preprint arXiv:2006.06091 (2020).

[2] ISO/TC 22 Road vehicles, ISO/TR 4804:2020: Road Vehicles — Safety and Cybersecurity for Automated Driving Systems — Design, Verification and Validation, first ed., International Organization for Standardization, 2020. URL: https://www.iso.org/standard/80363.html.

[3] A. Vivekanandan, N. Maier, J. M. Zöllner, Plausibility verification for 3d object detectors using energy-based optimization, in: Computer Vision–ECCV 2022 Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part I, Springer, 2023, pp. 602–616.

[4] J. Nitsch, M. Itkina, R. Senanayake, J. Nieto, M. Schmidt, R. Siegwart, M. J. Kochenderfer, C. Cadena, Out-of-distribution detection for automotive perception, in: 2021 IEEE International Intelligent Transportation Systems Conference (ITSC), IEEE, 2021, pp. 2938–2943.

[5] M. Keser, A. Savkin, F. Tombari, Content disentanglement for semantically consistent synthetic-to-real domain adaptation, in: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2021, pp. 3844–3849.

[6] Q. M. Rahman, P. Corke, F. Dayoub, Run-time monitoring of machine learning for robotic perception: A survey of emerging trends, IEEE Access 9 (2021) 20067–20075.

[7] T. Ponn, T. Kröger, F. Diermeyer, Identification and explanation of challenging conditions for camera-based object detection of automated vehicles, Sensors 20 (2020) 3699.

[8] M. Yin, S. Li, C. Song, M. S. Asif, A. K. Roy-Chowdhury, S. V. Krishnamurthy, Adc: Adversarial attacks against object detection that evade context consistency checks, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 3278–3287.

[9] É. Zablocki, H. Ben-Younes, P. Pérez, M. Cord, Explainability of deep vision-based autonomous driving systems: Review and challenges, International Journal of Computer Vision 130 (2022) 2425–2452.

[10] Y. Zhang, P. Tiňo, A. Leonardis, K. Tang, A survey on neural network interpretability, IEEE Transactions on Emerging Topics in Computational Intelligence 5 (2021) 726–742.

[11] S. Taxonomy, Definitions for terms related to driving automation systems for on-road motor vehicles, SAE: Warrendale, PA, USA 3016 (2018).

[12] E. Yurtsever, J. Lambert, A. Carballo, K. Takeda, A survey of autonomous driving: Common practices and emerging technologies, IEEE access 8 (2020) 58443–58469.

[13] Y. Tian, K. Pei, S. Jana, B. Ray, Deeptest: Automated testing of deep-neural-network-driven autonomous cars, in: Proceedings of the 40th international conference on software engineering, 2018, pp. 303–314.

[14] S. Tolmeijer, M. Christen, S. Kandul, M. Kneer, A. Bernstein, Capable but amoral? comparing ai and human expert collaboration in ethical decision making, in: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, 2022, pp. 1–17.

[15] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 618–626.

[16] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, W. Samek, On pixel-wise

explanations for non-linear classifier decisions by layer-wise relevance propagation, PloS one 10 (2015) e0130140.

[17] Y.-C. Liu, Y.-A. Hsieh, M.-H. Chen, C.-H. H. Yang, J. Tegner, Y.-C. J. Tsai, Interpretable self-attention temporal reasoning for driving behavior understanding, in: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2020, pp. 2338–2342.

[18] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in neural information processing systems 30 (2017).

[19] P. Jacob, É. Zablocki, H. Ben-Younes, M. Chen, P. Pérez, M. Cord, Steex: steering counterfactual explanations with semantics, in: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XII, Springer, 2022, pp. 387–403.

[20] M. Harradon, J. Druce, B. Ruttenberg, Causal learning and explanation of deep neural networks via autoencoded activations, arXiv preprint arXiv:1802.00541 (2018).

[21] C. Aytekin, Neural networks are decision trees, arXiv preprint arXiv:2210.05189 (2022).

[22] K. Pei, Y. Cao, J. Yang, S. Jana, Deepxplore: Automated whitebox testing of deep learning systems, in: proceedings of the 26th Symposium on Operating Systems Principles, 2017, pp. 1–18.

[23] M. Bansal, A. Krizhevsky, A. Ogale, Chauffeurnet: Learning to drive by imitating the best and synthesizing the worst, arXiv preprint arXiv:1812.03079 (2018).

[24] D. Feng, A. Harakeh, S. L. Waslander, K. Dietmayer, A review and comparative study on probabilistic object detection in autonomous driving, IEEE Transactions on Intelligent Transportation Systems 23 (2022) 9961–9980. doi:10.1109/TITS.2021.3096854.

[25] J. Choi, D. Chun, H. Kim, H.-J. Lee, Gaussian yolov3: An accurate and fast object detector using localization uncertainty for autonomous driving, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 502–511.

[26] N. Vallez, A. Velasco-Mata, O. Deniz, Deep autoencoder for false positive reduction in handgun detection, Neural Computing and Applications 33 (2021) 5885–5895.

[27] M. Keser, G. Schwalbe, A. Nowzad, A. Knoll, Interpretable model-agnostic plausibility verification for 2d object detectors using domain-invariant concept bottleneck models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 3890–3899.

[28] M. S. Khesbak, Depth camera and laser sensors plausibility evaluation for small size obstacle detection, in: 2021 18th International Multi-Conference on Systems, Signals & Devices (SSD), IEEE, 2021, pp. 625–631.

[29] F. Geissler, A. Unnervik, M. Paulitsch, A plausibility-based fault detection method for high-level fusion perception systems, IEEE Open Journal of Intelligent Transportation Systems 1 (2020) 176–186.

[30] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, P. Liang, Concept bottleneck models, in: International Conference on Machine Learning, PMLR, 2020, pp. 5338–5348.

[31] B. Wang, L. Li, Y. Nakashima, H. Nagahara, Learning bottleneck concepts in image classification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 10962–10971.