Safety and Robustness for Deep Neural Networks: An Automotive Use Case*

Davide Bacciu[1], Antonio Carta[1][0000−0002−0003−2323], Claudio Gallicchio[1][0000−0002−6692−2564], and Christoph Schmittner[2]

[1] Department of Computer Science, University of Pisa, Pisa PI, Italy
{davide.bacciu, antonio.carta, claudio.gallicchio}@unipi.it
[2] Austrian Institute of Technology, Vienna
christoph.schmittner@ait.ac.at

Abstract. Current automotive safety standards are cautious when it comes to utilizing deep neural networks in safety-critical scenarios due to concerns regarding robustness to noise, domain drift, and uncertainty quantification. In this paper, we propose a scenario where a neural network adjusts the automated driving style to reduce user stress. In this scenario, only certain actions are safety-critical, allowing for greater control over the model's behavior. To demonstrate how safety can be adressed, we propose a mechanism based on robustness quantification and a fallback plan. This approach enables the model to minimize user stress in safe conditions while avoiding unsafe actions in uncertain scenarios. By exploring this use case, we hope to inspire discussions around identifying safety-critical scenarios and approaches where neural networks can be safely utilized. We see this also as a potential contribution to the development of new standards and best practices for the usage of AI in safety-critical scenarios. The work done here is a result of the TEACHING project, an European research project around the safe, secure and trustworthy usage of AI.

Keywords: recurrent neural networks · adversarial robustness · human-in-the-loop · automotive · dependability

1 Introduction

The rapid advancements in artificial intelligence (AI) and machine learning (ML) have enabled the development of sophisticated deep neural networks (DNNs) that can perform complex tasks with great accuracy. However, the use of DNNs in safety-critical applications, such as autonomous driving, is still a topic of debate. While DNNs have shown great promise in improving the safety and efficiency of automotive systems, their usage in safety-critical scenarios remains a concern due to the difficulty in ensuring robustness to noise, domain drift, and uncertainty quantification.

* This research was supported by TEACHING, a project funded by the EU Horizon 2020 research and innovation programme under GA n. 871385.

Currently, automotive safety standards recommend avoiding the use of DNNs in safety-critical scenarios, given the potential risks associated with their usage. We argue that there are scenarios where DNNs can be used safely, provided that their actions are restricted and monitored. In this paper, we present a scenario where a DNN is used to adapt the driving style to minimize the user's stress. In this scenario, only certain actions of the DNN are safety-critical, allowing for greater control over the model's behavior.

To ensure the safety of the DNN, we propose a mechanism based on robustness quantification and a fallback plan. This ensures that the DNN can minimize user stress in safe conditions while avoiding unsafe actions in uncertain scenarios. By exploring this use case, we hope to inspire discussions around identifying safety-critical scenarios where DNNs can be safely utilized. This could influence the development of new standards and best practices for the usage of AI in safety-critical automotive applications.

The objectives of the paper are:

1. We show that current standards suggest to avoid the use of DNNs (Technology class III) in safety-critical scenarios;
2. we argue that DNN can be used in scenarios where there is a minimal safety risk (Usage Level C) and the model can be easily restricted to a subset of safe actions;
3. we describe a Level C use case which uses recurrent networks where we quantify the adversarial robustness and ensure safety by restricting the actions in any unsafe setting.

The work presented here was done in TEACHING, short for "A computing Toolkit for building Efficient Autonomous appliCations leveraging Humanistic INtelliGence,". TEACHING is an European Union-funded research project aimed at designing a comprehensive computing platform and associated software toolkit. This platform and toolkit are intended to support the development and deployment of autonomous, adaptive, and dependable Cyber-Physical Systems of Systems (CPSoS) applications. The project focuses on enabling these applications to leverage sustainable human feedback to drive, optimize, and personalize the provisioning of their services.

The TEACHING project revolves around four key concepts:

- Distributed Edge-oriented and Federated Computational Environment: The project aims to create an integrated computational environment that seamlessly combines heterogeneous resources, including specialized edge devices, general-purpose nodes, and cloud resources. One important aspect is the utilization of edge devices equipped with specialized hardware to execute AI, cybersecurity, and dependability components of autonomous applications.
- Runtime Dependability Assurance of CPSoS: TEACHING focuses on developing methods and tools to ensure runtime dependability assurance for CPSoS. This includes the establishment of systematic engineering processes for designing both conventional and AI-based runtime adaptive systems. These approaches will be applied in both cloud and edge environments to guarantee

continuous assurance throughout the software life cycle, incorporating AI methodologies tailored towards a cognitive security framework.
- Software-level Abstraction of the Computing System: The project aims to realize a software-level abstraction of the computing system, enabling the easy and coordinated deployment of different application components onto suitable CPSoS resources. This concept also involves orchestrating application components to optimize resource efficiency, minimize energy consumption, and meet the dependability requirements of the application.
- Synergistic Human-CPSoS Cooperation: TEACHING emphasizes the collaboration between humans and CPSoS in the spirit of Humanistic Intelligence. It explores AI methodologies and continuous monitoring of human physiological, emotional, and cognitive (PEC) states to facilitate applications with unprecedented levels of autonomy and flexibility. This collaboration maintains the required dependability standards for safety-critical systems operating with humans in the loop.

By focusing on these four main concepts, the TEACHING project aims to advance the development and deployment of efficient and dependable autonomous applications within CPSoS, while leveraging humanistic intelligence and ensuring safety-critical operations.

In Section 2, we will describe the current automotive standards with a focus on recommendations regarding safety and AI. Section 3.4 provides an overview of the deep neural network literature relevant for our system. Both sections cover the relevant state of the art, from a standardization and technical point of view. In Section 3 we describe our scenario in more detail, discuss the challenges associated with using DNNs in safety-critical scenarios, and propose our mechanism for ensuring the safety of the DNN. Finally, we discuss the potential implications of our work and how it could contribute to the development of safer and more efficient automotive systems (Section 4).

2 Automotive Standards for Dependability

Functional safety standards for automotive applications (road vehicles) are in the focus of ISO TC22 SC32, WG08. This group has created the basic functional safety standard for road vehicles, ISO 26262 (Ed.2, not considering nominal performance issues; now preparing for Ed 3, also including new technologies like AI and SotiF-related issues) and ISO 21448 SotiF, Safety of the intended Functionality, considering uncertainties of environment and functional insufficiencies impacting even vehicles safety fulfilling ISO 26262 functional safety requirements). Automated Driving Systems safety is handled in TR 4804, which is published as “Safety and cybersecurity for automated driving systems — Design, verification and validation” and will be superseded by TS 5083, “Safety for automated driving systems”. The following figure shows the interrelationships between these standards (Figure 1). These standards focus on dependability for road vehicles considering also the impact of AI systems integrated and nominal performance issues caused by functional insufficiencies. In the TEACHING project, the safe

functioning of the ADS is assumed. The ASIL safety integrity levels include also as part of the assessment the controllability by the driver, The role of the human is also part of some of the trustworthiness standards which development recently started, e.g., like ISO/IEC PWI 18966: Artificial Intelligence (AI) — (human) Oversight of AI systems, and other human-related or ethics and governance related standards. In automated driving, when, particularly in case of an identified failure during operation or if the vehicle leaves the ODD (Operational Design Domain) for which it was designed (due to an unexpected mismatch of environment or situational conditions), the ADS has to request a takeover to the driver or passenger, the human reaction will become safety-relevant. How this has to be taken into account when assessing the functional safety of the AI system monitoring human health and awareness (alertness) conditions will be discussed during our presentation of the scenario.

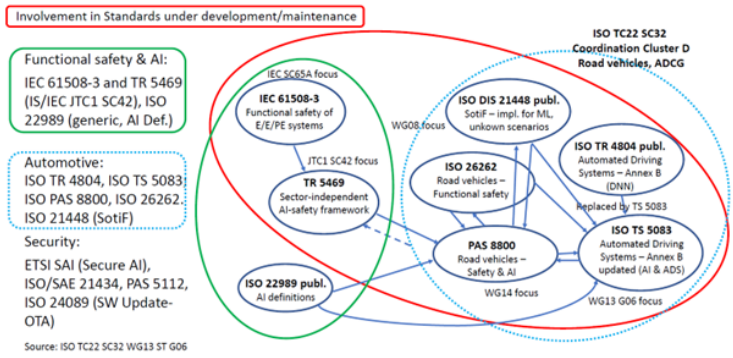

Fig. 1: Automotive standardization landscape on functional safety and AI (source: ISO TC22/SC32 WG13).

2.1 Current AI Standards Recommendations on Robustness and Functional Safety

ISO and IEC created a Joint Technical Committee (JTC1) aiming for harmonized standardization in the AI-sector. The ISO/IEC JTC1 SC42 Standardization Committee for "Artificial Intelligence" has published already 17 ISO/IEC standards under its own responsibility and is developing (in different stages of development) 30 new standards. The most important WG is WG03, Trustworthiness. ISO/IEC JTC1 SC42, WG03, "Trustworthiness", started together in cooperation with the maintenance team of IEC 61508-3, the basic functional safety standard, SW part, to develop TR 5469 "Functional safety and AI systems". The approach taken, in short, was to classify AI technology classes and usage classes. The following figure tries to map these and provide recommendations:

AI Technology Classes

Class I developed and reviewed using existing functional safety methods and standards.

Class II cannot be fully developed and reviewed using existing functional safety methods and standards, but it is still possible to identify a set of available methods and techniques satisfying the properties (e.g., additional V&V). to achieve the necessary risk reduction

Class III cannot be developed and reviewed using existing functional safety methods and international functional safety standards and it is also not possible to identify a set of available methods and techniques satisfying the functional safety properties.

AI Application and Usage Classes

A1 Used in safety relevant E/E/PE system and automated decision making possible.

A2 Used in safety relevant E/E/PE system and no automated decision making (e.g., for uncritical diagnostics). B1: Used to develop safety relevant E/E/PE systems (offline support tool). Automated decision making of developed function is possible.

B2 Used to develop safety relevant E/E/PE systems (offline support tool). No automated decision making of the developed function is possible.

C AI technology is not part of a safety function in the E/E/PE system. Has potential indirect impact on safety (e.g., increase demand placed on a safety system).

D AI technology is not part of a functional safety function in the E/E/PE system, but can have indirect impact on the function (e.g., increase demand placed on a safety system).

3 Automotive Use Case

In this section, we describe the use case. It is a use case where a module of the system continuously predicts the stress of the user. The prediction is made via a pre-trained Recurrent Neural Network [7], which is a deep neural network designed to process time series.

In this use case, the system continuously monitors the stress of the users. In safe condition, it tries to minimize the user's stress by controlling the driver profile. For example, some users may become more stressed because the current driving style is too slow, while others may be stressed due to the high acceleration and speeds. An adaptive model learn the user's preference and adaptive the driving profile for them automatically.

The user's stress is also monitored to control the activation of some driving profiles. In some settings, the profile may only be allowed if the user is ready to take action in case of emergency (ISO TS 5083). If the user is deemed too stressed, some of the driving profiles may be temporarily disabled until the stress

<table>
<tr><th>AI Technology Class => AI application and usage level</th><th>AI technology Class I</th><th>AI technology Class II</th><th>AI technology Class III</th></tr>
<tr><td>Usage Level A1 (1)</td><td rowspan="5">Application of risk reduction concepts of existing functional safety International Standards possible</td><td>Appropriate set of requirements (3)</td><td rowspan="5">At the time of writing this document no appropriate set of properties with related methods and techniques is known to achieve sufficiently reduction of risk</td></tr>
<tr><td>Usage Level A2 (1)</td><td>Appropriate set of requirements (3)</td></tr>
<tr><td>Usage Level B1 (1)</td><td>Appropriate set of requirements (3)</td></tr>
<tr><td>Usage Level B2 (1)</td><td>Appropriate set of requirements (3)</td></tr>
<tr><td>Usage Level C (1)</td><td>Appropriate set of requirements (3)</td></tr>
<tr><td>Usage Level D (2)</td><td colspan="3">No specific functional safety requirements for AI technology, but application of risk reduction concepts of existing functional safety International Standards</td></tr>
<tr><td colspan="4">1 Static (offline) (during development) teaching or learning only
2 Dynamic (online) teaching or learning possible
3 The appropriate set of requirements for each usage level can be established by application of risk reduction concepts of existing functional safety International Standards and additional consideration of Clauses 8, 9, 10 and 11 of this document. Examples are provided in Annex B. Defining detailed requirements for each usage level is beyond the Scope of this document.</td></tr>
</table>

Fig. 2: RECOMMENDATIONS FOR USAGE OF AI TECHNOLOGY CLASSES IN CERTAIN USAGE LEVELS. (Source: DTR 5469, which is already in an advanced stage).

level is reduced. Notice that the stress is monitored only for the purpose of the automated driving system. The user can always disable the system and take over the control of the car. This is in accordance with the German Ethics guidelines, which state that the driver should always be allowed to take over the automated systems.

3.1 Use Case Description

In the scenario, the user is a passenger in a car driven by an autonomous driving system. The driving system supports different driving styles, such as a relaxed mode and a sport mode. We assume that the relax mode is fully autonomous. In contrast, the sport mode may have restrictions, such as geofencing some areas where the user must take the wheel, or allowing usage in some settings only if the user is awake, alert, and unstressed. The system monitors the passenger stress level and selects the driving mode that maximizes the user's comfort and minimizes stress with an adaptive model. As a result, there is a continuous interaction between the human passenger and the autonomous system.

Overall, the system is made of four components that interact in a closed loop:

Passenger The passenger is part of the system because its stress is a result of the surrounding environment, such as the car temperature, its speed, or the traffic. The passenger also produces outputs, its physiological data, that is used by the rest of the system to inform the autonomous decisions, possibly resulting in a change in the driving style. In case fitness and alertness

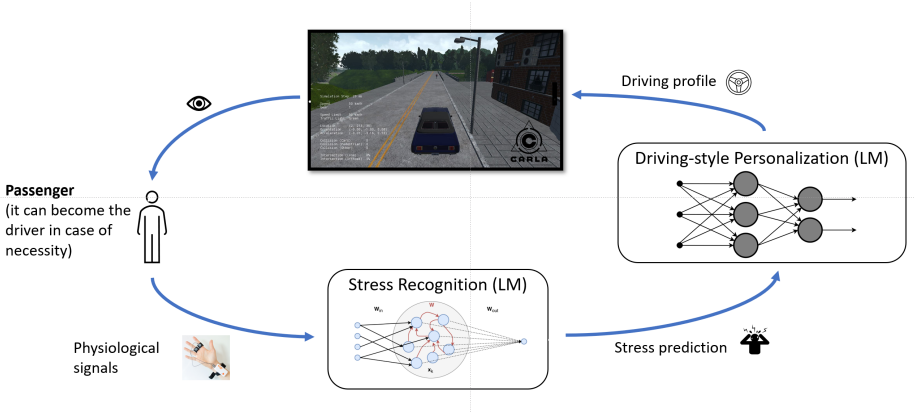

Fig. 3: Stress recognition use case.

are required (see take-over case as explained before) this data may be also safety-relevant.

Stress prediction The physiological data of the passenger, such as electrodermal activity and heart rate, are continuously monitored via sensors. These signals are fed as input to a *stress prediction module*, a Recurrent Neural Network (RNN) module that predicts the *stress of the passenger* given its physiological data. The output of this module is a time series of the stress predictions.

Driving style personalization The driving style personalization is a module that takes as input the stress predictions and the environmental data (state of the car and input from its sensors) and determines whether the driving style must be changed or not. The driving style model is a deep neural network.

Autonomous driving system Finally, a change in the autonomous driving system changes the state of the car (e.g. its speed) and as a result affects the stress of the user, closing the interaction loop.

A diagram of the entire system is shown in Figure 3.

3.2 Usage Level C and Deep Neural Networks

An initial analysis of the use case shows that the application class (driver monitoring, but not controlling the automated vehicle) is Usage Level D, which is uncritical from the viewpoint of the automated vehicle's behaviour and safety. In ISO TS 5083, "Automated Driving Systems," an evolving standard, there is a subgroup E10 ("Post deployment phase") that addresses the deployment and operation phase. In this subgroup, it is stated that in the event of a failure of the Automated Driving System (ADS) or when the Operational Design Domain (ODD) for which the vehicle was designed is exceeded, there may be a requirement to include the driver, passenger, or remote operator interaction. This scenario is

comparable to the situation of an airline passenger seated next to the emergency exit. In the event of an emergency, this passenger is required to take certain actions. Therefore, it is necessary to inquire if the passenger is physically capable of fulfilling this responsibility. If the passenger is unable to do so, an alternative passenger will need to occupy that seat. In this case, a risk of wrong detection of drivers/passengers fitness and alertness because of a high-stress level may have a safety impact. This case has to be studied separately, being of Usage Level C. In this case, appropriate requirements have to be considered and risk reduction measures taken. This may include strong signaling if within a defined short timeframe the driver does not react properly by taking over respectively the vehicle changes to situation-based degraded motion modes as discussed later.

If the autonomous driving system would be able to drive everywhere without limitations, this would be a Usage Level D use case. A mistake in the stress prediction or driving style personalization models may result in a higher stress for the user but it will not cause safety-critical issues, assuming that the autonomous driving system is a safe component.

However, due to law and safety regulations, there may be many requirements that would limit the usage of the autonomous system and therefore result in a Usage Level C use case, which requires a more attentive design in the two predictive models.

For example, there are some situations where the passenger or user of a vehicle or transportation system must be prepared to take action in case of an emergency. In these instances, it is essential that the passenger remains alert and ready to assume control of the vehicle or perform certain actions, such as opening emergency doors, if necessary. Suppose that the self-driving system is fully autonomous when driving at slower speeds, but may require the user's help in certain situations when driving at faster speeds. For instance, a sport driving mode may be available in certain areas only if the user remains alert, relaxed, and firmly grasping the steering wheel, prepared to take over control of the vehicle if an emergency arises.

In this scenario, the stress and driving style modules are part of a safety-critical system because the stress prediction module must determine if the user is too stressed and therefore the sport mode should be disabled. Conversely, the driving style module must never allow a driving style that is forbidden in the current setting.

This is a Usage Level C because activating the sport mode even if the user is stressed will increase the demand on the safety system. The same is valid for the cases mentioned before when alertness is required (TS 5083, post-deployment phase): fitness to react properly is endangered if the stress level had become too high and no proper action had been taken.

3.3 Stress Module Robustness

In this section, we focus on the problem of assessing the robustness of the stress predictor. As a robustness measure, we would like to quantify the robustness of the model to perturbations of the input specifically crafted to break it. These are

the worst-case perturbations and can be measures via adversarial robustness[12] methods.

The stress prediction model takes as input a sequence of sensor measurements $x(1), \ldots, x(t)$, where $\mathbf{x}(t) \in \mathbb{R}^{N_X} \in \mathbb{R}^{N_H}$, where we denote as $\mathbf{x}^t$ the current measurement. The chosen model for the stress predictor is a Recurrent Neural Network (RNN), which is a model that keeps an internatl state $\mathbf{h}(t) \in \mathbb{R}^{N_H}$ and updates it at each timestep as

$$\mathbf{h}(t) = \tanh\left(\mathbf{W}_{in}\mathbf{x}(t) + \hat{\mathbf{W}}\mathbf{h}(t-1)\right), \tag{1}$$

where $\mathbf{W}_{in} \in \mathbb{R}^{N_H \times N_X}$ is the input-to-recurrent parameter matrix, and $\hat{\mathbf{W}} \in \mathbb{R}^{N_H \times N_H}$ the recurrent matrix. We omit biases from the previous and the following equations to simplify the notation. The stress predictor outputs $\mathbf{y}(t) \in \mathbb{R}^{N_Y}$ with a linear layer on top of the hidden state as

$$\mathbf{y}(t) = \mathbf{W}\mathbf{h}(t), \tag{2}$$

where $\mathbf{W} \in \mathbb{R}^{N_Y \times N_H}$ is a hidden-to-output matrix. The network can be trained end-to-end via backpropagation-through-time. Alternatively, we can initialize the RNN parameters separately and train only the final classifier, which can be trained with a closed-form equation. This approach is known as reservoir computing[10]. An interesting consequence of partial training is that we could initialize the RNN parameters to improve its adversarial robustness. This an open question and a promising research direction.

Adversarial Robustness Quantification There are several methods to computed adversarial robustness bounds, such as POPQORN [12] and CertRNN [6]. Given a reference dataset of sequences $\mathbf{X} = \{\mathbf{x_0}, \ldots, \mathbf{x_n}\}$, a model F and its predictions on the reference dataset $\mathbf{Y} = \{\mathbf{y_0}, \ldots, \mathbf{y_n}\}$ and an l_p ball with radius ϵ, robustness quantification methods provide an upper and lower bound for the RNN's output for each sequence in $\mathbf{X}$ when subject to noise with radius ϵ. Therefore, given a target accuracy, we can find the maximum perturbation ϵ that achieves the target accuracy. We can use methods such as POPQORN to quantify the robustness of our trained model before deployment. This functionality is implemented in a learning module of the TEACHING platform [2]. According to ISO/IEC TR 24029-1:2021 definition, POPQORN is an empirical method.

The resulting method requires three parameters, each of which is critical for safety:

trained model The architecture of the model, its hyperparameters, training algorithm, and training data all affect the robustness of the method.

reference dataset The robustness quantification is an empirical method that relies on a curated dataset. Particular care must be taken to ensure that the dataset is as extensive as possible.

required accuracy Any amount of noise will result in a accuracy drop. Therefore, there must be a minimal required accuracy that must be guaranteed by the model even for the largest amount of noise.

Automotive standards and best practices should inform the creation of the dataset. The largest amount of operating conditions, such as types of roads, countries, weather conditions, must be represented in the dataset to ensure robustness in any given setting. The reference dataset is a test dataset used only for evaluation. To guarantee a fair evaluation, it must not be used during any step of the model training, such as the hyperparameter selection or preprocessing (e.g. computing normalization constants), or feature selection. If the reference dataset is used in any capacity, the resulting robustness bounds will be an overestimate of the actual robustness on new data.

Finally, every application must define a target accuracy. Given that we use a machine learning model, we cannot expect perfect accuracy, and we expect the accuracy for large perturbations to be even lower. The application must guarantee safety even when using an imperfect model, up to a certain target accuracy. The models can also be made more robust by ensembling different models.

The robustness quantification assessess whether the system accuracy is adequate and robust, and therefore deemed safe for each specific operating condition. For example, a system may be certified only in some specific settings, such as specific city zones, weather conditions, or road types. The system may fail to certify in specific conditions either due to lack of reference data, target accuracy, or target robustness. The fallback system ensures safety even in these scenarios.

Fallback system An interesting property of our use case is that only some of the choices are safety-critical (i.e., the sport mode under some specific conditions). As a result, it is always possible to make the model safer by excluding the safety-critical choices from the possible actions. This results effectively in a conversion of our Level C use case into a Level D, where no choice would result in harm to the user. A fallback mechanism can be used, where the safety-critical choices are possible only for the samples that are deemed close enough to the regions of interest where the network reaches the target accuracy. In case of uncertainty, the safety-critical choices can always be disabled to ensure the safety of the system even when an unsafe model is used.

Overall, the robustness quantification and fallback system work together to ensure safety. If the system detects a known and safe operating condition it enables all the drving modes, reverting to safe choices in unknown and unsafe scenarios. Detection of unsafe conditions can be measure offline via robustness quantification and implemented online by explicitly enabling the safe conditions with simple signals, e.g. via geofencing. Be default, an unknown operating condition is deemed unsafe, restricting the usage to the safe driving modes.

3.4 Related Work on Adversarial Robustness for Deep Neural Networks

It is well known that deep neural networks are susceptible to adversarial attacks [4]. An adversarial example is an input that has been modified to fool the DNN to make an incorrect classification by adding a small amount of well-crafted noise.

To the human eye, an adversarial example is often indistinguishable from the original data.

Adversarial attacks are a practical issue that strongly limits the deployment of DNN in autonomous vehicles. [13] shows that it is possible to add a sticker to an object, such as a road signals, such that the object will be consistently misclassified by the DNN classifier. There are also universal attacks which are able to make the DNN fail on all possible objects. For example, [13] shows that you can place a sticker over a camera to misclassify all the objects of a certain class. Such attacks are known as universal perturbations.

[1] identifies the problem of obfuscated gradients

Adversarial training consists in the training of a network with the original and adversarial examples during training. By training on adversarial examples directly, the model can become more robust to adversarial noise. [3] provides a taxonomy and a review of existing methods.

While many work propose defenses to improve adversarial robustness, most are quickly broken by new attacks or due to weaknesses in their defense evaluation [4].

[5] shows that unlabeled data can be used to significantly improve adversarial robustness in semisupervised learning.

While most work on adversarial examples focuses on images, our scenario uses Recurrent Neural Networks (RNNs) to process time series. POPQORN is a certification methods that allow to estimate robustness bounds for RNN models such as the Long Short-Term Memory (LSTM) [11]. More recently, CertRNN proposed a general framework for certifying the robustness of RNNs, providing an exact formula for the computation of bounding planes [6]. The computation results in a tighter bound that outperforms POPQORN.

Research on adversarial robustness is still very active, and it is an open question whether it is possible to fully fix robustness issues. [9] establishes fundamental limits on the robustness of some classifiers in terms of a distinguishability measure between the classes. Unlike previous works, [15] claims that it is possible to train robust and accurate models. The paper shows that regular adversarial examples leave the manifold, even though on-manifold adversarial examples exist, and they correspond to generalization errors. As a result, increasing the accuracy will increase the robustness and the two objectives are not necessarily contradicting goals. [8] proposes contrastive learning through the lens of robustness enhancement and proposes AdvCL, an adversarial contrastive pretraining framework which enhances cross-task robustness transferability. [14] introduce a novel regularizer that encourages the loss to behave linearly in the vicinity of the training data, thereby penalizing gradient obfuscation while encouraging robustness

4 Conclusion

In this paper, we describe an automotive use case with a human in the loop, as studied by the TEACHING project. This is a usage level C use case that can be solved using recurrent neural networks. However, current standards prevent

the usage of deep neural networks for level C use cases, except when appropriate mitigation measures are taken. We argue that our use case provides an example where neural networks can be applied, and encourage more discussion on the topic. In particular, technical solutions need to be developed to address both the robustness quantification, uncertainty estimation, and the development of fallback mechanisms. Extremely conservative fallback mechanisms can already be used today, but there is a need for more effective solutions. On the other hand, regulatory bodies and standards commitees must explore the consequences of the usage of neural networks in level C use cases, with a particular focus on the specification of robustness and accuracy targets, the curation of reference datasets, and best practices to develop effective fallback mechanisms.

References

1. Athalye, A., Carlini, N., Wagner, D.: Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples (Jul 2018). https://doi.org/10.48550/arXiv.1802.00420
2. Bacciu, D., Akarmazyan, S., Armengaud, E., Bacco, M., Bravos, G., Calandra, C., Carlini, E., Carta, A., Cassarà, P., Coppola, M., Davalas, C., Dazzi, P., Degennaro, M.C., Di Sarli, D., Dobaj, J., Gallicchio, C., Girbal, S., Gotta, A., Groppo, R., Lomonaco, V., Macher, G., Mazzei, D., Mencagli, G., Michail, D., Micheli, A., Peroglio, R., Petroni, S., Potenza, R., Pourdanesh, F., Sardianos, C., Tserpes, K., Tagliabó, F., Valtl, J., Varlamis, I., Veledar, O.: TEACHING - Trustworthy autonomous cyber-physical applications through human-centred intelligence. In: 2021 IEEE International Conference on Omni-Layer Intelligent Systems (COINS). pp. 1–6 (Aug 2021). https://doi.org/10.1109/COINS51742.2021.9524099
3. Bai, T., Luo, J., Zhao, J., Wen, B., Wang, Q.: Recent Advances in Adversarial Training for Adversarial Robustness (Apr 2021). https://doi.org/10.48550/arXiv.2102.01356
4. Carlini, N., Athalye, A., Papernot, N., Brendel, W., Rauber, J., Tsipras, D., Goodfellow, I., Madry, A., Kurakin, A.: On Evaluating Adversarial Robustness (Feb 2019). https://doi.org/10.48550/arXiv.1902.06705
5. Carmon, Y., Raghunathan, A., Schmidt, L., Duchi, J.C., Liang, P.S.: Unlabeled Data Improves Adversarial Robustness. In: Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019)
6. Du, T., Ji, S., Shen, L., Zhang, Y., Li, J., Shi, J., Fang, C., Yin, J., Beyah, R., Wang, T.: Cert-RNN: Towards Certifying the Robustness of Recurrent Neural Networks. South Korea p. 19 (2021)
7. Elman, J.L.: Finding structure in time. Cognitive science **14**(2), 179–211 (1990)
8. Fan, L., Liu, S., Chen, P.Y., Zhang, G., Gan, C.: When does Contrastive Learning Preserve Adversarial Robustness from Pretraining to Finetuning? In: Advances in Neural Information Processing Systems. vol. 34, pp. 21480–21492. Curran Associates, Inc. (2021)
9. Fawzi, A., Fawzi, O., Frossard, P.: Fundamental limits on adversarial robustness
10. Gallicchio, C., Micheli, A., Pedrelli, L.: Deep reservoir computing: A critical experimental analysis. Neurocomputing **268**, 87–99 (Dec 2017). https://doi.org/10.1016/J.NEUCOM.2016.12.089
11. Hochreiter, S., Schmidhuber, J.: Long Short-Term Memory. Neural Computation **9**(8), 1–32 (1997). https://doi.org/10.1144/GSL.MEM.1999.018.01.02

12. Ko, C.Y., Lyu, Z., Weng, L., Daniel, L., Wong, N., Lin, D.: POPQORN: Quantifying Robustness of Recurrent Neural Networks. In: Proceedings of the 36th International Conference on Machine Learning. pp. 3468–3477. PMLR (May 2019)
13. Li, J., Schmidt, F.R., Kolter, J.Z.: Adversarial camera stickers: A physical camera-based attack on deep learning systems. arXiv:1904.00759 [cs, stat] (Jun 2019)
14. Qin, C., Martens, J., Gowal, S., Krishnan, D., Dvijotham, K., Fawzi, A., De, S., Stanforth, R., Kohli, P.: Adversarial Robustness through Local Linearization. In: Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019)
15. Stutz, D., Hein, M., Schiele, B.: Disentangling Adversarial Robustness and Generalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6976–6987 (2019)