Confidence Arguments for Evidence of Performance in Machine Learning for Highly Automated Driving Functions

Simon Burton[1,3], Lydia Gauerhof[2], Bibhuti Bhusan Sethy[2], Ibrahim Habli[3], and Richard Hawkins[3]

[1] Systems Engineering Vehicle, Robert Bosch GmbH, Schwieberdinger Str. 76, 71636 Ludwigsburg, Germany
[2] Corporate Research, Robert Bosch GmbH, Robert-Bosch-Campus 1, 71272 Renningen, Germany
{Simon.Burton, Lydia.Gauerhof, BibhutiBhusan.Sethy}@de.bosch.com
[3] Assuring Autonomy International Programme, The University of York, York, U.K.
{Ibrahim.Habli, Richard.Hawkins}@york.ac.uk

Abstract. Due to their ability to efficiently process unstructured and highly dimensional input data, machine learning algorithms are being applied to perception tasks for highly automated driving functions. The consequences of failures and insufficiencies in such algorithms are severe and a convincing assurance case that the algorithms meet certain safety requirements is therefore required. However, the task of demonstrating the performance of such algorithms is non-trivial, and as yet, no consensus has formed regarding an appropriate set of verification measures. This paper provides a framework for reasoning about the contribution of performance evidence to the assurance case for machine learning in an automated driving context and applies the evaluation criteria to a pedestrian recognition case study.

Keywords: Highly Automated Driving · Machine Learning · Safety Assurance.

1 Introduction

Highly Automated Driving (HAD) has the potential to radically decrease the number of road accidents as well as introducing significant convenience and ecological benefits. At the same time, HAD functions are themselves safety-critical and must therefore be demonstrated to meet strict safety criteria before their release for use on public roads. Existing safety standards such as ISO 26262 [3] define prerequisites that must be fulfilled to minimise the risk of hazards caused by random hardware and systematic failures in the electrical/electronic systems. Due to the complexity of the systems and inherent uncertainty in the operating environment, HAD systems also require an increased focus on demonstrating that hazards are not caused by inherent restrictions in the sensors, actuators or

decision logic. ISO PAS 21448 [1] addresses the "Safety of the Intended Functionality" by considering such effects. However, this standard is currently focused on Level 1 to 2 [4] driver assistance systems rather than Level 3 to 5 HAD systems which include higher levels of autonomy and for which machine learning is seen as a key enabling technology.

Machine learning algorithms and in particular Deep Neural Networks (DNNs) [15] are being applied to the task of providing an accurate perception for highly automated driving functions. One of the challenges caused by applying machine learning methods to these tasks is that a precise specification of the required behaviour is often not possible. Indeed, it is the very fact that the machine learning functions are able to infer the target function without a detailed specification, based on the presented training data that makes them so appealing. The lack of a precise specification combined with the unpredictable and opaque nature of the algorithms introduce high degrees of uncertainty into the safety assurance process.

This paper is organised as follows: A generic safety case pattern for arguing the performance of machine learning models previously proposed by the authors is summarised in Section 2. This is then used to derive a model for reasoning about the contribution of evidence to this assurance case pattern in Section 3 and used to formulate a corresponding confidence argument approach. In Section 4, the confidence argument approach is then applied to techniques that have been developed for verifying DNN-based perception functions for highly automated driving. Feature map sensitivity analysis is also used to provide counter-evidence for the confidence argument. The paper closes with a discussion of the need for a more rigorous approach to developing and proposing performance evaluation methods within a safety context and proposes future work in this area.

2 Safety Case Patterns for Machine Learning

In order to support the claim that the Machine Learning Model (MLM) meets its performance requirements, it is important to understand the causes of such insufficiencies. As interest in machine learning safety has grown, a number of authors [6], [25], [26] have investigated different causes of performance limitations in machine learning functions. Some examples applicable to HAD are described below:

- **Distributional shift:** Critical or ambiguous situations, within which the system must react in a predictably safe manner, may occur rarely or may be so dangerous that they are not well represented in the training data. It must be argued that the training data contains an appropriate distribution of all classes of critical situations and object classes or that the selected training leads to an appropriate level of generalisation. In addition, the system should continue to perform safely even if the operational environment differs from the training environment over time [6].
- **Robustness deficits of the trained function:** An adversarial perturbation [16], [19], [20] is an input sample that is similar (at least to the human

eye) to other samples but that leads to a completely different categorisation with a high confidence value. It has been shown that such examples can be automatically generated and used to trick the network. The challenge, therefore, is to ensure that the machine learning algorithms focus on those properties of the inputs relevant to the target function without becoming distracted by irrelevant features. In other words, act within the same hierarchical dimensions as the target function [18].

– **Differences between the training and execution platforms:** When using machine learning to represent a function that is embedded as part of a wider system, the input to the neural network will have typically been processed by a number of elements already [25], such as lenses, image filters and buffering mechanisms. These elements may vary between the training and target execution environments leading to the trained function becoming dependent on hidden features of the training environment not relevant in the target system.

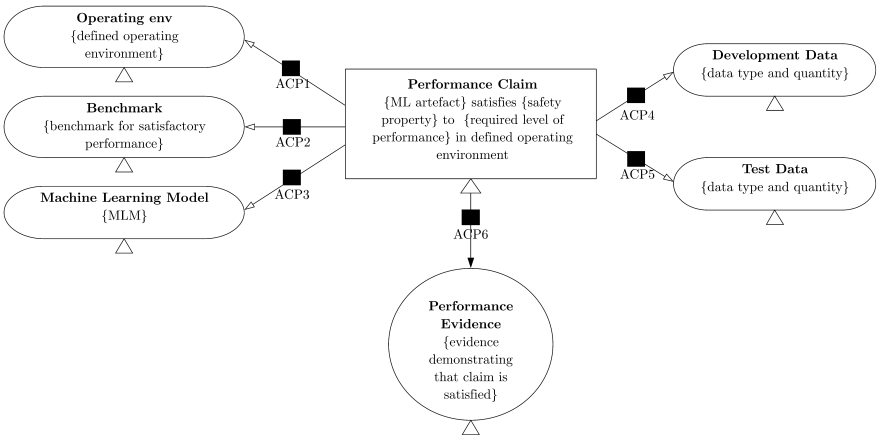


Fig. 1: Safety case pattern for machine learning model

Previous work by the authors as well as others have introduced concepts of applying assurance case structures to arguing the performance of an MLM within a safety-critical context [8], [17], [22]. Figure 1 describes a generic assurance case pattern for arguing the safety properties of a machine learning function (derived from the description in [23] using GSN [2]). This assurance case pattern is centered on discharging the claim that the MLM fulfills its safety properties (defined by benchmarks) to a required level of performance in a defined operation environment.

A contract-based approach to specifying safety properties of the MLM was proposed in [8], by which the MLM is specified as a component within its sys-

tem context and defined by a set of assumptions on its operating environments under which certain safety guarantees (for example formulated as benchmark performance requirements) must hold. These performance requirements could include definitions of accuracy and failure rates to be achieved by the function. This allows for the assurance case for the MLM to focus on the safety-relevant properties of the trained function whilst the validity of the assumptions and appropriateness of the guarantees are discharged as part of a system level assurance activity.

In contrast to classical software-based approaches, existing safety standards do not define a set of accepted methods for evaluating the performance of machine learning in a safety critical context. Therefore during assessment and homologation, any proposed assurance case will inevitably lead to questions regarding the strength of argument presented and the relevance of the presented supporting evidence. Assurance Claim Points (ACPs) [12], indicated by the black squares in the pattern, are used to represent points in the argument where further assurance is required through the provision of a more detailed confidence argument. The confidence in the assurance case is therefore achieved by supporting the claims within the following ACPs. ACP1 and ACP2 must be supported by arguments that consider the overall system context [11], whilst ACPs 3...6 are specific to the machine learning function. These confidence arguments can be then used to aid the certification process, especially where accepted best practice has yet to be defined.

- **ACP1**: Argument that the assumptions made on the operational design domain as well as on the interfaces to other technical components within the system are valid.
- **ACP2**: Argument that the benchmark performance requirements allocated to the guarantees of the safety contract for the MLM are sufficient to fulfill the overall system safety requirements.
- **ACP3**: Argument that the adopted training process and the choice of model and hyperparameters lead to a function that fulfills its requirements.
- **ACP4**: Argument that the training data are sufficient to lead to a MLM that fulfills its performance requirements.
- **ACP5**: Argument that the test data that is used is sufficient to support the performance claim.
- **ACP6**: Argument that the performance evidence generated from the test data is sufficient to support the performance claim.

3 Confidence Arguments for Performance Evidence

In this section we develop the concept of performance evidence confidence. This confidence argument will then support the claim that the provided evidence sufficiently supports the performance claim. In order to derive the set of conditions to be discharged by the confidence arguments we introduce a number of definitions, which will be defined in the set of equations below. These definitions

are used here to illustrate the relationships between elements of the assurance case in order to stimulate a discussion regarding under what conditions these relationships hold true and as such the authors do not intend the definitions to necessarily form a mathematically complete model. In general, the performance claim can be formulated as a simple equivalence between the specified behaviour of the system and actual behaviour.

$$\forall i \in I.M(i) = T(i) \tag{1}$$

Where i is a sample from the actual input domain I, M represents the trained model and T the specification (or ground truth) for a given input. In other words, for all possible inputs of the input domain, the implementation provides the same result as the specification. The application of the design-by-contract approach allows us to formulate a more restrictive form of equivalence that constrains the input space that fulfills the set of assumptions and limits the properties of interest to those formulated in the guarantees. This can be formulated as follows:

$$\forall i \in I.A(i) \Rightarrow G(i, M(i)) \tag{2}$$

In other words, for all possible inputs in the domain that fulfill the set of explicitly specified assumptions A, the implementation provides a result that meets the safety guarantees G for the given inputs.

Equation 2 can now also be used to define the concept of *Contract Performance* by defining the conditional probability of a safety contract being fulfilled over the set of inputs that fulfill its assumptions A. The assurance case claim that the machine learning function fulfills its guarantees G with a conditional probability (ρ) can therefore be defined as follows:

$$\forall i \in I.\rho(G(i, M(i))|A(i)) > ContractPerformance \tag{3}$$

The confidence argument that a given evidence leads to an adequate assessment of the actual performance of the machine learning function can therefore be couched in terms of the relationship between the measurement provided by the evidence and the actual contract performance as described in Equation 3. In order to perform this comparison, it is necessary to define a measurement value threshold (*MeasurementTarget*) provided by the evidence E that, if reached, is postulated to imply that the *ContractPerformance* target is met. This allows for the following definition of *EvidenceContribution* to the safety case performance claim:

$$\begin{gathered} \forall i \in I, \exists S \subseteq I, \forall j \in S. \\ (A(j) \wedge (E(S) > MeasurementTarget)) \Rightarrow \\ \rho(G(i, M(i))|A(i)) > ContractPerformance \end{gathered} \tag{4}$$

Where E is a function that takes as input a set of samples (S) from the input domain that meet the defined set of assumptions and returns a quantifiable measure that can be compared against a target value. In its simplest form, E could represent simply tests on selected inputs and return the proportion of

tests that passed. The testing problem could thus be formulated as finding some minimum subset S of the input domain to use as test data such that whenever the test results pass a pre-defined target, then the performance over the entire valid input space meets the contract performance.

E could also represent a more indirect measure that is used to infer the performance of the machine learning function such as the robustness towards adversarial perturbations. The definition of *EvidenceContribution* can also be extended to combine a number of different evidences which must all fulfill their measurement targets in order to imply that the *ContractPerformance* is met, thus allowing for combining of a mixture of techniques and measurements into E and *MeasurementTarget*.

The definition of *EvidenceContribution* allows us to identify several claims that need to be made as part of the confidence arguments ACP5 and ACP6 as described in Section 2. ACP5 can be strengthened by providing evidence to support the claims:

- The sample set used to provide performance evidence is capable of detecting faults in the machine learning function that would lead to a violation of performance requirements.
- The sample set is representative of the input domain and the application of the performance evaluation on this sample set leads to a representative indication of the measurement target for the entire domain.

ACP6 can be strengthened by providing evidence to support the claims:

- There is a demonstrable correlation between the *MeasurementTarget* and the *ContractPerformance*.
- The measurements based on the sample set can be extrapolated to provide an indication of the expected performance for the entire input domain even in the case of root unknown causes of insufficiencies (in ISO PAS 21448 defined as unknown triggering events).

4 Case Study

In this section we apply the assurance case structure described above to the pedestrian recognition case study introduced in [10] and demonstrate how arguments regarding typical performance evaluation techniques can be strengthened or refuted. The performance requirements of the function used for the case study can be summarised as follows:

- Pedestrians of width X pixels and height Y pixels are classified.
- Pedestrians are detected if C% of the person is occluded.
- There are less than FP% of false positive classifications per frame.
- There are less than FN% of false negative classifications per frame.
- Vertical deviation from the ground truth is less than V pixels.
- Horizontal deviation from the ground truth is less than H pixels.

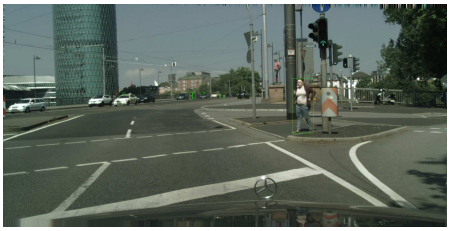
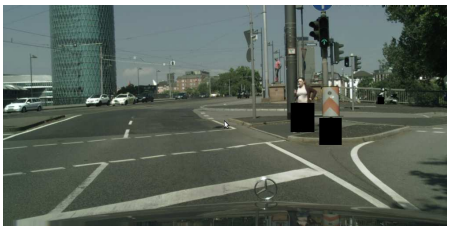


Image example from CityPersons including the ground truth as green bounding boxes [27]



Lower part of predictions are masked. The woman is predicted correctly, while the road sign is a false positive.

Fig. 3: Image example from CityPersons [27] with ground truth and partly masked

For the purpose of our case study we focus on the requirement that pedestrians should be detected even if certain portions of the person are occluded. This is based on the assumption that in the operating environment pedestrians may be partially occluded by objects such as street furniture or baby strollers. A typical approach to collecting performance evidence for such requirements would be to ensure that the test data contained examples of occluded and non-occluded persons. This would lead to the following instantiation of Equation 4 to describe the relationship between the testing approach and the performance claim:

$$\begin{gathered} \forall i \in I, \exists Testset \in\subseteq I, \forall j \in Testset. \\ (A_{occlusion}(j) \wedge TestsPassed(Testset) > TestBenchmark) \Rightarrow \\ \rho(G(i, M(i))|A_{occlusion}(i)) > ContractPerformance \end{gathered} \tag{5}$$

Where $A_{occlusion}$ describes assumptions on the input data including that pedestrians may be occluded and $TestsPassed$ is the evidence function that returns the proportion of tests passed based on the sample set $Testset$ which also includes occluded persons. The Guarantee function G here represents the combination of performance requirements described above where $ContractPerformance$ defines the required level of conditional probability that the performance requirements are met in the field (overall target failure rate). $TestBenchmark$ represents a target proportion of the tests that should pass as part of the release process for the Machine Learning function. In reality, a set of assumptions and evidence measures would be combined to evaluate the performance requirements, not just relying on assumptions regarding occluded persons. In order to evaluate the confidence arguments related to "Adequacy of the sample set to discover faults" and "Representativeness of the sample set", we applied an experimental approach to investigate the correlation between occlusion of parts of the pedestrian and activations within the DNN. In [9] a visualization technique was introduced that gives insight into intermediate feature layers of a DNN. This method demonstrates which input pattern of the image causes the activation of a particular feature map. In our experiment, we use the same diagnostic method to trace the feature map activities back to the input pixel space [9]. For this

Layer Name	Output Channel No.	Activation for human lower part, unmasked Image	Activation for human lower part masked Image
fire5/expand3x3	108	80.68%	25.0%
maxpool5	236	84.81%	25.0%
fire6/squeeze1x1	5	81.62%	14.67%
fire6/expand3x3	72	84.63%	5.0%
fire6/concat	264	84.63%	25.0%
fire7/expand3x3	74	80.29%	19.67%
fire7/concat	266	80.29%	17.67%
fire8/concat	20	85.3%	18.33%
fire8/concat	165	80.52%	21.67%
fire9/squeeze1x1	49	80.95%	20.5%

Table 1: Sensitivity analysis of feature maps for unmasked images and masked lower part of pedestrians. Chosen layers are mainly activated for lower part of pedestrians.

purpose, we trained a Squeeznet [14] on CityPersons [27]. We then evaluated the resulting activation map not only manually, but also statistically.

For our experiment, we apply the diagnostic method to search for the feature maps which are activated by the lower part of the body by investigating the activation map [9]. After identifying the relevant feature maps, we verify this dependency through statistical evaluation. We mask the lower 50% of all detected pedestrians from the CityPersons data set, as shown in Figure 3, and compare the activations against unmasked images. If the feature map is activated, the mean pixel value of the lower part of the bounding box in the activation map $ActiveMap_{lowBB}$ is higher than the mean pixel value of the total activation map $ActiveMap_{total}$. Eq.: 6 describes the activation of the feature map:

$$\frac{\sum_{p=1}^{\#ActiveMap_{lowBB}} ActiveMap_{lowBB}[p]}{\#ActiveMap_{lowBB}} > \frac{\sum_{d=1}^{\#ActiveMap_{total}} ActiveMap_{total}[d]}{\#ActiveMap_{total}} \quad (6)$$

The sensitivity analysis in Table 1 is conducted on the Munich test data set of CityPersons[27] with 383 images. The layers are mainly activated, when the lower part of the detected pedestrian is visible (third column). However, they are less activated, when the lower part is masked (forth column in Table 1). This analysis confirms the activation of the feature map is particularly sensitive to the visibility of the lower part of body. Consequently, we provide evidence that the relevant feature map for detecting the lower body is not activated, when the lower body is masked. Evaluation only on the prediction would not reveal what caused each prediction. This leads us to reassess the potential of the test data sets at detecting faults related to occlusion of different body parts. Furthermore, this sensitivity analysis can be now extended to other feature maps to find additional weaknesses in the DNN and identify suitable counter-measures. These could include the retraining of particular layers or of the whole DNN.

Confidence Argument	Description
Adequacy to discover faults	Test data sets can give an overall evaluation of performance. However, they do not necessarily reveal specific systematic performance issues (such as undue focus on lower body when detecting pedestrians). In addition, this technique is not well suited to uncover robustness issues of the trained function.
Measurement relevance	Faults discovered when applying the test data set directly indicate weaknesses in the trained model with respect to realistic input data. However, issues regarding differences between the target environment and environment used to collect training and test data must also be addressed.
Sample set is representative	The test data set is likely to contain similar biases caused by scalable oversight, distributional shift to that of the training data. In addition, the sample set may be representative of the distribution of features in the input domain, however this may not guarantee the detection of critical rarely occurring corner cases.
Extrapolation of results	The performance targets that can be argued are limited with respect to the size and distribution of the data set and are not focused towards particular causes of insufficiencies.

Table 2: Summary of confidence claims for test data sets

5 Evaluation of Performance Evidence Approaches

Based on the confidence argument structure described in Section 3, we can now assess performance evaluation techniques regarding their contribution to the performance claim that a particular MLM fulfills its performance criteria. Table 2 summarises an evaluation of confidence case elements for testing based on test data sets including some insights provided by the case study described above. This analysis highlights several of the weaknesses associated with test data driven verification of machine learning functions and demonstrates the need for strong supporting evidence in the confidence argument to ensure that issues such as fault coverage and sample set representativeness are addressed.

A key weakness associated with such techniques is their apparent inability to detect robustness deficits that may not be related to feature dimensions directly relevant to the properties of the operating environment of interest.

Next, we assess confidence arguments for techniques that analyse the robustness of a trained function against adversarial perturbations, and in particular those that make use of introspection techniques. In Table 3 we investigate the concept outlined in [13]. In this approach, the robustness of the trained network is verified by demonstrating that regions within the input space exhibit a similarity within the activation network such that misclassifications in the case of adversarial inputs cannot occur, where the adversarial inputs may be deliberately manipulated or due to other effects such as sensor noise.

Confidence Argument	**Description**
Adequacy to discover faults	The analysis focuses on faults caused by adversarial perturbations that exploit robustness deficits in the trained function. A fault model is defined in the form of perturbations against which the trained function shall be robust and the region within the input space in which the perturbations are deemed relevant.
Measurement relevance	The technique relies on a number of assumptions to allow for a tractable analysis. These include the relevance of features within (hidden) layers of the network and the size of the regions to be analysed (amount of perturbation). The correlation between the parameters of the analysis and their relevance to the overall performance in relation to the actual *ContractPerformance* is unclear.
Sample set is representative	The technique performs an exhaustive search of particular regions of the neural network. The analysis is sound for a given bounded input space region. However the analysis is only performed for specific images. Therefore the results will depend greatly on the selection of the images as a starting point for the analysis.
Extrapolation of results	Due to the uncertainty regarding the relevance of the performance measurement and the representativeness of the sample set, a method for extrapolating the results of the performance evaluation across the entire input domain was unclear and is likely to rely on a number of specific assumptions and constraints.

Table 3: Summary of confidence claims for analysing robustness against adversarial perturbations [13]

6 Summary and future work

This paper has shown that existing approaches to evaluating the performance of machine learning in the context of safety-related automated driving functions provide evidence of only limited value for a safety assurance case. This is admittedly a non-trivial task and as yet no industry consensus or standards exist regarding which combination of techniques should be applied for the performance evaluation of such functions. An approach was provided for constructing confidence arguments for performance evaluation techniques which could be used in future work to demonstrate their contribution to the assurance case and the conditions under which the contributions are valid. The approach was used to evaluate a pedestrian recognition function and sensitivity analysis of feature maps was used to highlight weaknesses in the trained function and also to reflect on the contribution of typical performance evaluation techniques.

The evaluations described in Section 5 highlight the fact that each individual performance evaluation technique is limited according to a certain set of constraints and assumptions. By better understanding these, for example through the use of techniques such as sensitivity analysis of feature maps (as described in our experiment), introspection methods [21, 5], fault injection [24], mutation testing [7], a combination of evidence may be found that provides a convincing argument that the performance requirements are met. Explicitly evaluating

the machine learning approach and its performance evaluation measure against the set of claims defined in the assurance claim points leads to a greater level of confidence that the performance requirements have been met. This in turn can provide additional support for safety assessment and certification activities, especially in the absence of accepted best practice and standards.

Future work will focus on deepening the understanding of insufficiencies in the MLMs by performing sensitivity analysis for a wider range of features whilst providing stronger confidence arguments for any proposed evidence to support the performance claim. The authors also propose the use of confidence arguments in future standardisation efforts in order to better motivate the contribution of particular evaluation techniques, or to provide a framework by which the use of any particular combination of techniques can be justified for a particular system context.

References

1. ISO/PRF PAS 21448: Road vehicles - safety of the intended functionality. Tech. rep., International Standards Organisation (ISO), Geneva (2011)
2. Goal structuring notation community standard version 2. Tech. rep., Assurance Case Working Group (ACWG), https://scsc.uk/r141B:1?t=1, accessed on 04/05/2019 (2018)
3. ISO 26262: Road vehicles - functional safety, second edition. Tech. rep., International Standards Organisation (ISO), Geneva (2018)
4. SAE j3016: Surface vehicle recommended practice, (r) taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. Tech. rep., SAE International, Geneva (2018)
5. Alsallakh, B., Jourabloo, A., Ye, M., Liu, X., Ren, L.: Do convolutional neural networks learn class hierarchy? CoRR **abs/1710.06501** (2017), http://arxiv.org/abs/1710.06501
6. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., Mané, D.: Concrete problems in ai safety. arXiv preprint arXiv:1606.06565 (2016)
7. Baker, R., Habli, I.: An empirical evaluation of mutation testing for improving the test quality of safety-critical software. IEEE Transactions on Software Engineering **39**(6), 787–805 (2012)
8. Burton, S., Gauerhof, L., Heinzemann, C.: Making the case for safety of machine learning in highly automated driving. In: International Conference on Computer Safety, Reliability, and Security. pp. 5–16. Springer (2017)
9. Chollet, F.: Deep Learning with Python. Manning Publications Co., Greenwich, CT, USA, 1st edn. (2017), chapter: 5.4.1. Visualizing intermediate activations
10. Gauerhof, L., Munk, P., Burton, S.: Structuring validation targets of a machine learning function applied to automated driving. In: International Conference on Computer Safety, Reliability, and Security. pp. 45–58. Springer (2018)
11. Hawkins, R., Habli, I., Kelly, T.: The principles of software safety assurance. In: 31st International System Safety Conference (2013)
12. Hawkins, R., Kelly, T., Knight, J., Graydon, P.: A new approach to creating clear safety arguments. In: Advances in systems safety, pp. 3–23. Springer (2011)
13. Huang, X., Kwiatkowska, M., Wang, S., Wu, M.: Safety verification of deep neural networks. In: International Conference on Computer Aided Verification. pp. 3–29. Springer (2017)

14. Iandola, F.N., Han, S., Moskewicz, M.W., Ashraf, K., Dally, W.J., Keutzer, K.: SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. arXiv e-prints arXiv:1602.07360 (Feb 2016)
15. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105 (2012)
16. Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial examples in the physical world. arXiv preprint arXiv:1607.02533 (2016)
17. Kurd, Z., Kelly, T.: Establishing safety criteria for artificial neural networks. In: International Conference on Knowledge-Based and Intelligent Information and Engineering Systems. pp. 163–169. Springer (2003)
18. Lin, H.W., Tegmark, M., Rolnick, D.: Why does deep and cheap learning work so well? Journal of Statistical Physics **168**(6), 1223–1247 (2017)
19. Metzen, J.H., Genewein, T., Fischer, V., Bischoff, B.: On detecting adversarial perturbations. arXiv preprint arXiv:1702.04267 (2017)
20. Nguyen, A., Yosinski, J., Clune, J.: Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 427–436 (2015)
21. Nguyen, A.M., Yosinski, J., Clune, J.: Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks. CoRR **abs/1602.03616** (2016), http://arxiv.org/abs/1602.03616
22. Picardi, C., Habli, I.: Perspectives on assurance case development for retinal disease diagnosis using deep learning. In: 17th Conference on Artificial Intelligence in Medicine. Springer (2019)
23. Picardi, C., Hawkins, R., Paterson, C., Habli, I.: A pattern for arguing the assurance of machine learning in medical diagnosis systems. In: International Conference on Computer Safety, Reliability, and Security. Springer (2019)
24. Schorn, C., Guntoro, A., Ascheid, G.: Efficient on-line error detection and mitigation for deep neural network accelerators. In: International Conference on Computer Safety, Reliability, and Security. pp. 205–219. Springer (2018)
25. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.F., Dennison, D.: Hidden technical debt in machine learning systems. In: Advances in neural information processing systems. pp. 2503–2511 (2015)
26. Varshney, K.R.: Engineering safety in machine learning. In: 2016 Information Theory and Applications Workshop (ITA). pp. 1–5. IEEE (2016)
27. Zhang, S., Benenson, R., Schiele, B.: CityPersons: A Diverse Dataset for Pedestrian Detection. arXiv e-prints arXiv:1702.05693 (Feb 2017)