EasyChair Preprint

№ 6847

A Methodology to Use AI for Automotive Safety Using Model Explainability

Srikanth Kaniyanoor Srinivasan, V Krishna and
Harsha Vardhan Sahoo

October 16, 2021

A Methodology to use AI for Automotive Safety using Model Explainability

Srinivasan, Srikanth Kaniyanoor
Intel Corporation
Bengaluru, India
srikanth.kaniyanoor.srinivasan@intel.com

V, Krishna
Intel Corporation
Bengaluru, India
krishna.v@intel.com

Sahoo, Harsha Vardhan
Intel Corporation
Bengaluru, India
harsha.vardhan.sahoo@intel.com

***Abstract* — Workload consolidation, use of powerful processors and the advancement in Artificial Intelligence (AI) have enabled deployment of complex algorithms on high compute processors. Over time, the ability to derive effective algorithms using AI has increased dramatically paving their use in safety critical application. However, their adequacy from the perspective of safety certification remains controversial. In this paper we explain a methodology on how to use AI for safety critical applications using explainable AI (XAI) techniques. The approach is explained taking a use case of traffic light detection (TLD) as a part of active safety implementation targeting ASIL B. Recommendations from recent standards like SOTIF and UL 4600 are incorporated to augment the safety case.**

Keywords — XAI, AI, DL, FuSa, ISO 26262, SOTIF, UL 4600.

I. Background. Motivation and Objectve

Dramatic success of machine learning and deep learning algorithms together with the advancement of processors has triggered an accelerated growth of complex autonomous systems incorporating AI models in diverse applications. These autonomous systems are developed with an objective to build intelligent agents that perceive, learn, decide, and act on their own. Most of the DL models are designed to make accurate predictions for a given static dataset. The models are black box by nature because they are created directly from static dataset which means that even humans who create the model might seldom understand how the variables are combined to obtain the target prediction. Therefore, their performance in real-life scenarios are not predictable and hence discouraged for use in safety critical applications [1].

Algorithms like linear or logistic regression and decision trees are easy to understand on how the decisions are made with built-in support for feature visualization and its importance. Deep learning models like CNNs, YOLO, SSD_MobileNet, etc are not easily interpretable or explainable due to the depth in terms of the number of layers and their ability to automatically learn features based on the input dataset. Due to the nature in which they learn and perceive complex features, they are termed as *black box models* thus, making explainability difficult. Unless a model can be interpreted or explained, it becomes difficult to deploy them for safety critical applications [2] [3].

The terms interpretability and explainability are often confused. In simple terms, interpretable models are those models where the cause and effect can be observed within the system. They are also referred to as *white box models*. On the other hand, explainable models are those where the internal working of a model can be explained in human terms. An explainable model can be designed to explain the working of the underlying DL model there by augmenting the trust in the DL model. Typically, classification problems using DL can be the right pick for explainability, whereas regression models and decision trees can be considered as interpretable models. Although the functional safety standards (FuSa) like ISO 26262 do not discourage use of DL models, they expect suitable justification and the methodology followed to implement the model. SOTIF and UL 4600 are emerging standards where guidelines are being formulated for use of AI in safety critical applications. In this paper we cherry pick recommendations from the upcoming standards and apply it by taking a real-life use case. The implementation flow is explained considering a use case where a DL model is used for traffic light detection (TLD) and warning system as a part of active safety function. The target ASIL considered for this use case is ASIL B.

Remaining part of this paper is organised into four sections:

- Section II provides the requirements and expectations of ISO 26262 standards for model-based development.
- Section III provides implementation details for the use case considered
- Section IV provides the results and discussion
- Section V summarises the findings and identifies the scope of future work.

II. Model based design using ISO 26262

ISO 26262-part 6 Annexure B gives detailed requirements for using model-based design in safety critical application. The key requirements are captured in *Table 1*

Table 1 - Key requirements from ISO 26262-6 Annexure B

Requirement	Proposed implementation
Models shall be developed using formal or semi formal notations	Model summary for the given AI model along with a textual description of the accuracies and the limitations of the model can be documented using semi-formal notations
Modelling guidelines shall be specified. Key features to be captured include comprehensibility, correct transformation, and execution of the model	One way to achieve this requirement is by using XAI where the explanations can help in documenting the comprehensibility of the model apart from the accuracy and precision of the model.
Software safety requirements	Can be adopted SOTIF standards where the training

	datasets derived based on the HARA can be used as requirements [4].
Development of the software architecture	Software architecture shall consider list of system limitations and counter measures. For example, a computer vision system might perform poorly if sufficient illumination is not present. Using sensor fusion to provide a weighted probability of the target class could overcome this limitation.
Software Unit testing as specified in section B.3.4 [5]	Since it is not easy to verify AI models at a unit level, the model can be verified using input images for various scenarios and XAI applied to obtain the explanations for the selected class. If the explanations are in line with what a human user would consider, then it can be marked as a PASS
Verification static/dynamic as specified in section B.3.4 [5]	Verification of AI model can be done with unseen datasets which were never used in training or testing as a part of unit test. Typical dataset taken from real world scenarios can serve as a verification set. If the model accuracies are greater than 90% for the sample selected, then we could consider the model as safe to use under the conditions documented in the safety manual.
Creation of a valid Safety case (adopted from UL 4600 standards	Design of a robust safety case which includes three elements viz, safety goal, an argument to satisfy the goal and evidence to verify the argument. UL 4600 aims at technology agnostic and goal-oriented safety case.

In section III, the implementation of the above requirements will be presented considering a use case where a DL model is used for detecting traffic lights and provides a visual/audio indication to the driver, with a target of meeting ASIL B.

III. MODELING OF TRAFFIC LIGHT DETECTION

Traffic light detection is an important requirement in autonomous navigation. The video feed can be obtained from a dashcam which can be processed using a DL model. This model detects presence of a traffic light in the scene and identifies which of the three lights are illuminated. In this paper a benchmark model for traffic light detection using ssd_mobilenet_v2 [6]. The model is trained using 150 images composed of 50 each of red, yellow, and green traffic light. The model is verified using 33 images. The goal of this benchmark model is to show a methodology of augmenting safety and not to discuss about the training strategy or data set selected for this model.

A. Safety goal and HARA analysis

In this section, we provide the safety goals for TLD use case. HARA analysis is performed considering UL 4600 standards where the safety case is goal oriented and technology agnostic. A detailed HARA analysis is performed to identify the risks and possible mitigation strategies.

Table 2 - Safety goal and HARA *analysis*

Hazard Description	Targeted ASIL	Safety Goal	Verification Requirements
TLD can fail to detect an illuminated traffic light.	ASIL B	The DL model shall be explainable using a human interpretable method.	The accuracy of the DL model is verified. The model is explained using a diverse redundancy.
TLD can incorrectly detect an illuminated traffic light causing a confusion for the driver.	ASIL B	TLD model shall provide an accuracy $\geqslant$ 90%	The accuracy of detecting a traffic light shall be greater than 90% along with the redundant model provideing an accuracy $\geqslant$ 70%. Note since the explainer represented using a mathematical model a lower accuracy.
TLD can fail to detect an illuminated traffic light due to background illumination	ASIL B	TLD model shall be trained with different background illumination.	Review of the dataset which contains all driving scenarios as specified in the SOTIF standards.

Based on the HARA analysis, TLD is trained based on the recommendations provided by SOTIF specification Annexure G [4]. The model is self-verified by looking at the training accuracies. As mentioned in [4], TLD model verification alone might be insufficient since it is difficult to ensure that the learning system has trained on the essential characteristics of the training data instead of coincidental correlations. The

model is verified using an explainable AI approach which is referred to as TLD_Explainer. A weighted average of TLD and TLD_Explainer can be considered for arriving at the final prediction. It can be mathematically expressed as

$$y_{target} = f(TLD, TLD - Explainer) \quad ---(1)$$

Weights for TLD and TLD_Explainer can be determined based on the background illumination as the model might perform differently during the day or night. The model can be fused with ambient light sensor which helps to determine the histogram thresholds.

IV. Results and Discussion

TLD model was implemented using SSD_MobileNet V2. Training data set consist of images obtained from different traffic intersections. The dataset is collected during different times of day as specified in Table F.1 in [4]. A distribution of 50% day, 35% night and 15% dusk dataset were considered. The training dataset was annotated into four classes vis. {Class A = Red light illuminated, Class B = Yellow light illuminated, Class C = Green light illuminated, Class D = No light illuminated}. The model is trained with a mean average precision 92.3%.

Since the model isn't explainable, the inferences obtained from the model are passed to an explainer to verify the model accuracy. The process steps are captured in Figure 1.

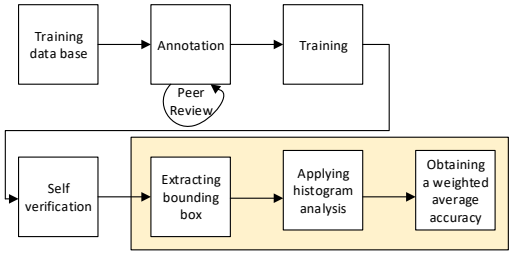

Figure 1 - Process Flow

A. Obtaining Inference explanations

TLD_Explainer helps in providing a justification for the prediction results obtained from TLD and can be used as a diverse method of recalculating the output. Furthermore TLD_Explainer is a simple mathematical model and can be easily expressed as a mathematical function. The process flow for implementing the diverse redundancy is shown in Figure 2.

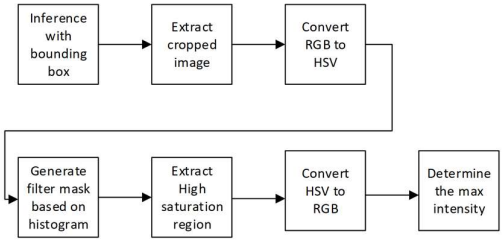

Figure 2 - Process Flow TLD_Explainer

The output of TLD as shown in Figure 3 is passed on to TLD_Explainer. The first step is to extract the ROI based on the bounding box information obtained from TLD. The ROI is converted from the RGB color space to HSV. This is shown in Figure 4. From the HSV image, saturation values are calculated, and a histogram is plotted to pick the filter mask that can be applied to extract out the significant region of interest as shown in Figure 5 and Figure 6

Once the filter mask is applied based on the value and saturation, the HSV image is converted back into RGB values. The highest value obtained from the RGB values indicate the target class for the given traffic light color.

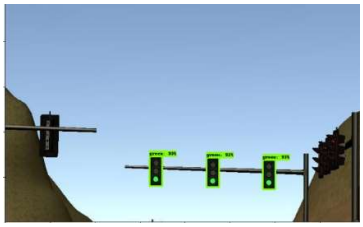

Figure 3 - Original scene image and the output of TLD model

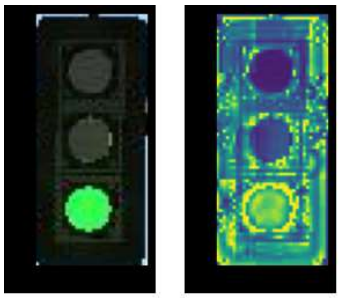

Figure 4 - Extraction of bounding box and color space conversion

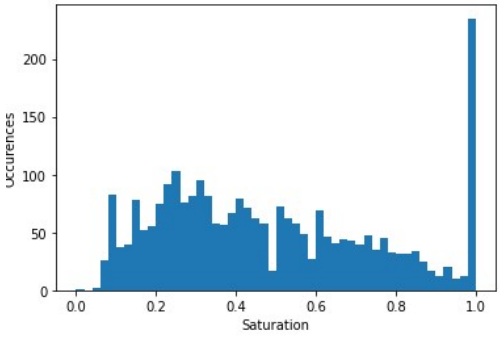

Figure 5 - Histogram

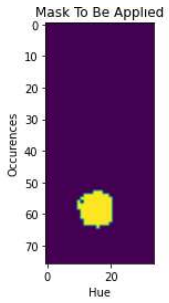

Figure 6 - Filter Mask

The final verdict is calculated by applying a weighted average of the predictions obtained from TLD and the TLD_Explainer. Since TLD_Explainer can be represented as a mathematical equation, a higher weightage is given to TLD_Explainer. In this paper, the weightage given to TLD is 0.75 and for TLD_Explainer is 1.25.

B. Results obtained for green light detection

The original image as show in Figure 3 was passed to TLD model and it identified three green traffic light with a detection probability of 92%, 93% and 90% respectively.
The output of this model was passed to the TLD_Explainer which predicted the light intensity as green with a 75% probability. Hence the final verdict that can be considered for the safety function is

$$\frac{(0.92 * 0.75 + 0.75 * 1.25)}{2} * 100 = 81\%$$

Since the verdict is greater than 70%, we can consider that the classification to belong to Class C Green light.
Based on the obtained confidence from both the models, the safety loop can be authorized to either take supervisory action or provide an indication to the driver. The decision by the safety loop can be defined as a function of the confidence levels obtained. For example, the safety loop can take a decision to slow down the vehicle in case the confidence from both the models is high and restrict itself to a warning indication otherwise.

The concept presented can be implemented on any ASIL/SIL certified processors like Intel® x6000FE. This processor implements a software lock step which can be used for obtaining the desired safety metric as per the standards. A reference implementation can be found in [7].

V. Conclusion and Future Work

The above work explains a methodology to use DL/ML techniques for augmenting safety for automotive applications. We demonstrated how the requirements mentioned in SOTIF, UL 4600 and ISO 26262 can be integrated to ensure incorporation of complex algorithms in safety critical applications without compromising the safety goals.

In future we propose to consider primary models, which have faster inferencing rates but poor explainability due to the underlying complexity of the model architecture. The outcome of these models can be passed to surrogate models which can be implemented using CNN and explained using existing XAI techniques like GradCAM [8] or LIME [9], thereby augmenting safety. This methodology can be applied for use cases like pedestrian detection, lane departure warning, obstacle detection, etc.

References

[1] C. Rudin and J. Radin, “Why Are We Using Black Box Models in AI When We Don’t Need To? A Lesson From An Explainable AI Competition,” *Harvard Data Sci. Rev.*, vol. 1, no. 2, 2019, doi: 10.1162/99608f92.5a8a3a3d.

[2] D. Gunning and D. Aha, “DARPA’s Explainable Artificial Intelligence (XAI) Program,” *AI Mag.*, vol. 40, no. 2, pp. 44–58, 2019, doi: 10.1609/aimag.v40i2.2850.

[3] Y. Lou, R. Caruana, and J. Gehrke, “Intelligible models for classification and regression,” 2012, doi: 10.1145/2339530.2339556.

[4] ISO, *ISO PAS 21448 - SOTIF standards*, 1st ed. Japan, 2017.

[5] ISO, *ISO 26262 part 6*, 2nd ed. 2016.

[6] Y. Takahashi, “Train a traffic light classifier using Tensorflow Object Detection API.” https://github.com/yuki678/driving-object-detection/blob/master/README.md.

[7] T. Wilkening, J. O. Krah, M. Salardi, and F. Heinzelmann, “Safety-Related High-Performance Motion Control based on a Quad-Core SoC,” in *PCIM Europe digital days 2021; International Exhibition and Conference for Power Electronics, Intelligent Motion, Renewable Energy and Energy Management*, 2021, pp. 1–8.

[8] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization,” *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, Oct. 2019, doi: 10.1007/s11263-019-01228-7.

[9] M. T. Ribeiro, S. Singh, and C. Guestrin, “‘Why should i trust you?’ Explaining the predictions of any classifier,” 2016, doi: 10.1145/2939672.2939778.