Safety monitoring of ML-based perception tasks of autonomous systems

Raul Sena Ferreira Jean Arlat Jeremie Guiochet Helene Waeselynck

LAAS-CNRS / University of Toulouse-UPS -- email: (rsenaferre, jean.arlat, jeremie.guiochet, helene.waeselynck) @laas.fr

Introduction

Machine learning (ML) provides no guarantee of safe operation in safety-critical systems such as autonomous vehicles. ML decisions are based on data that tends to represent a partial and imprecise knowledge of the environment. Such probabilistic models can output wrong decisions even with 99% of confidence, potentially leading to catastrophic consequences. Therefore, a fault tolerance mechanism, such as a safety monitor (SM), should be applied to guarantee the property correctness of these systems. However, applying an SM for ML components can be complex in terms of detection and reaction. Thus, aiming at dealing with this challenging task, this work presents a benchmark architecture for testing ML components with SM, and the current work for dealing with specific ML threats. We also highlight the main issues regarding monitoring ML in safety-critical environments.

Research challenges

For complex applications , designing an SM can be intractable due to the need to verify millions or even billions of parameters generated by the ML model. Therefore, this research tries to answer four research questions (RQ):

1. What type of ML threats can be detected at runtime?
2. How to monitor ML threats at runtime?
3. How to benchmark different runtime monitors?
4. How to intervene after the detection?

Benchmarking perception tasks of autonomous systems

For RQ1, we chose to focus on image classification tasks at runtime. There are several threats for this task at design or runtime as illustrated at Figure 1.

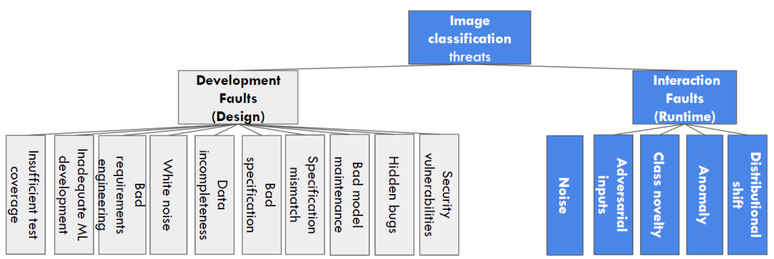

Figure 1:ML threats for image classification tasks.

For RQ2, we chose to build a combined approach using techniques capable of inspecting not just the observable parameters of the DNN that impact on its decision such as input features, neuron patterns, but also other properties of the system.

Regarding RQ3, this thesis compares different SM techniques by using an experimental framework based on the FARM[1] methodologyas illustrated in Figure 2.

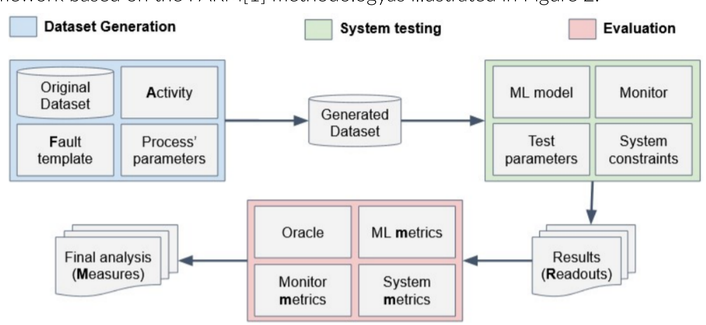

Figure 2:Benchmark architecture.

For RQ4, there are promising alternatives for reacting when a detection is made. For example, using a modified simplex architecture [5] with two controllers or synthesizing safety rules just for critical counterexamples [2].

Preliminary results for novelty class detection

- Six metrics: Mathews coefficient correlation (MCC), false positive rate (FPR), false negative rates (FNR), precision, recall and micro-f1.
- Three image datasets, varying between in-distribution data (ID), or out-of-distribution data (OOD): German Traffic Sign (GTRSB), CIFAR-10 and Belgium Traffic Sign (BTSC).
- Four SMs: three variants of outside-of-the-box [3] (OOB, OOB ISOMAP, OOB PCA), and out-of-distribution Image detector (ODIN) [4].

Table 1:Comparing data-based monitors for GTSRB as ID dataset, and BTSC as OOD dataset.

Method	MCC	FPR	FNR	Precision	Recall	Micro-F1
OOB	**0.21**	0.84	0.04	0.8	0.96	**0.73**
OOB ISOMAP	0.2	**0.72**	0.11	**0.81**	0.89	0.73
OOB PCA	0.04	0.86	0.11	0.78	0.89	0.68
ODIN	0.03	0.99	**0.0**	0.16	**1.0**	0.06

Table 2:Comparing data-based monitors for CIFAR-10 as ID dataset, and GTSRB as OOD dataset.

Method	MCC	FPR	FNR	Precision	Recall	Micro-F1
OOB	0.06	0.97	**0.0**	0.16	**1.0**	0.09
OOB ISOMAP	0.04	0.98	0.0	0.16	1.0	0.07
OOB PCA	0.17	0.8	0.03	0.2	0.97	0.33
ODIN	**0.23**	**0.61**	0.1	**0.24**	0.9	**0.52**

Table 3:Comparing data-based monitors for GTSRB as ID dataset, and CIFAR-10 as OOD dataset.

Method	MCC	FPR	FNR	Precision	Recall	Micro-F1
OOB	0.16	0.73	0.1	0.21	0.9	0.4
OOB ISOMAP	**0.19**	**0.06**	0.81	0.4	0.19	**0.8**
OOB PCA	-0.06	0.99	**0.02**	**0.64**	**0.98**	0.5
ODIN	-0.07	1.0	0.02	0.17	0.98	0.06

According to preliminary results, the performance of current monitors based on data are insufficient for monitoring these tasks.

3-year thesis work plan

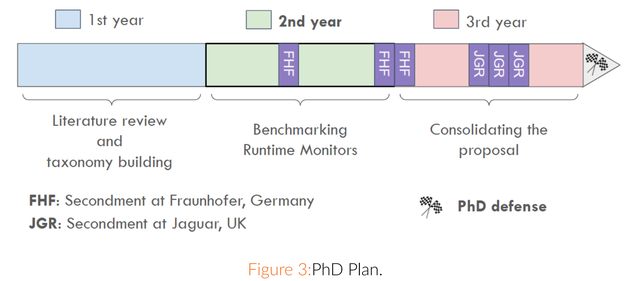

Figure 3:PhD Plan.

Acknowledgements

The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 812.788 (MSCA-ETN SAS). This publication reflects only the authors' view, exempting the European Union from any liability. Project website: http://etn-sas.eu/.

References

[1] Jean Arlat, Martine Aguera, Louis Amat, Yves Crouzet, J-C Fabre, J-C Laprie, Eliane Martins, and David Powell. Fault injection for dependability validation: A methodology and some applications. *IEEE Transactions on software engineering*, 16(2):166--182, 1990.

[2] Tommaso Dreossi, Alexandre Donzé, and Sanjit A Seshia. Compositional falsification of cyber-physical systems with machine learning components. *Journal of Automated Reasoning*, 63(4):1031--1053, 2019.

[3] Thomas A. Henzinger, Anna Lukina, and Christian Schilling. Outside the box: Abstraction-based monitoring of neural networks. In *ECAI*, volume 325 of *Frontiers in Artificial Intelligence and Applications, Santiago de Compostela, Spain*, pages 2433--2440. IOS Press, 2020.

[4] Yen-Chang Hsu, Yilin Shen, Hongxia Jin, and Zsolt Kira. Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10951--10960, 2020.

[5] Dung Phan, Nicola Paoletti, Radu Grosu, Nils Jansen, Scott A. Smolka, and Scott D. Stoller. Neural simplex architecture, 2019.