Evaluation and Hardening of Embedded AI Modules for Safety and Security in Critical Systems

The growing integration of Artificial Intelligence (AI) modules into safety-critical embedded systems (autonomous vehicles, drones, industrial and medical devices) raises major safety and security concerns. These modules, often based on deep neural networks, are sensitive to both accidental faults and intentional attacks [1]-[5] that can alter their decisions. Ensuring their robustness under real-world conditions is therefore essential for trustworthy deployment. Current approaches mainly focus on software-level adversarial robustness or high-level fault tolerance. However, few methodologies jointly consider physical disturbances, embedded constraints, and AI decision integrity in critical systems.

This PhD aims to develop a unified methodology for evaluating and improving the robustness of embedded AI modules against various real-world and physical disturbances, encompassing both safety-related faults and security-related attacks. The work will: (1) Identify, model, and reproduce representative perturbations that may cause abnormal or unsafe behavior. (2) Evaluate their effects on performance, safety, and security metrics. (3) Propose and validate mitigation and hardening techniques at the model, system, and learning levels.

The targeted application will concern multi-sensors based systems for autonomous vehicles embedding AI perception or decision modules. The experiments will rely on embedded platforms available at LCIS, including electromagnetic fault injection benches. The methodology will address both safety-related disturbances (accidental faults) and security-related threats (intentional perturbations), highlighting the differences between them in embedded AI systems.

Approach and Methodology

1. Definition of robustness metrics combining accuracy, integrity, latency, and safety.
2. Formalization of realistic fault and attack scenarios *at different system levels and integration* (real-world, sensors, preprocessing chain, AI module).
3. Implementation and validation of representative fault and attack scenarios, using a combination of simulation and physical experimentation on embedded platforms, to evaluate the robustness of AI modules and the effectiveness of the proposed hardening techniques.
4. Cross-layer analysis of robustness techniques, evaluating interactions and complementarities between mitigation methods for different classes of disturbances (to avoid conflicting protections or redundant efforts).
5. Design and validation of new efficient countermeasures across system levels, developing and experimentally assessing strategies under embedded constraints.

Expected Outcomes

- A methodology for robustness and fault-injection evaluation of embedded AI modules.
- Experimentally validated hardening strategies improving both safety and security.
- A benchmark and reproducible framework to support future studies for improving cross-layer robustness techniques.
- Design recommendations for safe and secure integration of AI in critical embedded systems.

PhD Student Profile:

- Master's in Embedded Systems
- Master's in Computer Science
- Master's in Microelectronics
- Master's in Cybersecurity
- Master’s in AI

Skills:

- Computer Architecture
- ML-based AI
- Prototyping and Simulation of Digital Systems

Location: Grenoble INP LCIS, Valence

Contacts: vincent.beroulle@lcis.grenoble-inp.fr
louis.morge-rollet@lcis.grenoble-inp.fr
valentin.egloff@lcis.grenoble-inp.fr

To apply for this position, please send the following documents to the individuals listed above:

- Your CV
- A letter of motivation (in French or English)
- A copy of your Master's transcript (M1 and M2)
- Letters of recommendation

References

[1] M. Dumont, K. Hector, P.-A. Moellic, J.-M. Dutertre, et S. Pontié, « Evaluation of Parameter-based Attacks against Embedded Neural Networks with Laser Injection », *International Conference on Computer Safety, Reliability, and Security* (2023), doi: 10.48550/ARXIV.2304.12876.

[2] V. Moskalenko, V. Kharchenko, et S. Semenov, « Model and Method for Providing Resilience to Resource-Constrained AI-System », Sensors, vol. 24, no 18, p. 5951, janv. 2024, doi: 10.3390/s24185951.

[3] A. Bosio, P. Bernardi, A. Ruospo and E. Sanchez, "A Reliability Analysis of a Deep Neural Network," 2019 IEEE Latin American Test Symposium (LATS), Santiago, Chile, 2019, pp. 1-6, doi: 10.1109/LATW.2019.8704548.

[4] P. Rech, "Artificial Neural Networks for Space and Safety-Critical Applications: Reliability Issues and Potential Solutions," in IEEE Transactions on Nuclear Science, vol. 71, no. 4, pp. 377-404, April 2024, doi: 10.1109/TNS.2024.3349956.

[5] S. Burel, A. Evans and L. Anghel, "Zero-Overhead Protection for CNN Weights," 2021 IEEE International Symposium on Defect and Fault Tolerance in VLSI and Nanotechnology Systems (DFT), Athens, Greece, 2021, pp. 1-6, doi: 10.1109/DFT52944.2021.9568363.