Resilience of deep learning applications: where we are and where we want to go

Cristiana Bolchini
Dip. Elettronica, Informazione e Bioingegneria
Politecnico di Milano
Milano, Italy
cristiana.bolchini@polimi.it

Alberto Bosio
Lyon Institute of Nanotechnology
Ecole Centrale de Lyon
Lyon, France
alberto.bosio@ec-lyon.fr

I. Introduction

Deep Learning (DL) [1] is currently one of the most intensively and widely used predictive models in the field of machine learning. DL has proven to give very good results for many complex tasks and applications, such as object recognition in images/videos, natural language processing, robotics, aerospace, smart healthcare, and autonomous driving. Nowadays, there is intense activity in designing custom Artificial Intelligence (AI) hardware accelerators to support the energy-hungry data movement, speed of computation, and memory resources that DL requires to realize its full potential [2]. Furthermore, there is an incentive to migrate AI from cloud to edge devices, i.e., Internet-of-Things devices, to address data confidentiality issues and bandwidth limitations, and also to alleviate the communication latency, especially for real-time safety-critical decisions, e.g., in autonomous driving.

The High-Level Expert Group on AI set up by the European Commission published in 2020 ethics guidelines for trustworthy use of AI systems [3]. The second requirement concerns the technical robustness and safety. We can directly cite from the guidelines the following: "A crucial requirement for achieving trustworthy AI systems is their dependability (the ability to deliver services that can justifiably be trusted) and **resilience** (robustness when facing changes). Technical robustness requires that AI systems are developed with a preventative approach to risks and that they behave reliably and as intended while minimising unintentional and unexpected harm as well as preventing it where possible".

Hardware-Accelerated Artificial Intelligence (HW-AI), similar to traditional computing hardware, is subject to hardware faults (HW faults) that can have several sources: variations in fabrication process parameters, fabrication process defects, latent defects, i.e., defects undetectable at time-zero post-fabrication testing that manifest themselves later in the field of application, silicon ageing, and Single Event Upsets stemming from ionization. All these HW faults can cause operational failures, potentially leading to important consequences, especially for safety-critical systems.

While HW-AI exhibits a certain resilience to HW faults, akin to the robustness found in biological neural networks, the effects of faults may be catastrophic and need be investigated and managed. The statistical behavior of neural network architectures, coupled with their abundant redundancy and overprovisioning, naturally endows them with a built-in tolerance for HW faults. During the learning process, HW-AI can circumvent to a large extent HW faults, however, this does not cover also all HW faults that occur after training, when the system is running. This vulnerability has the potential to affect inference, significantly impacting DL predictions and jeopardizing the functionality of the application. As a result, ensuring the reliability of HW-AI platforms becomes paramount, especially in safety-critical and mission-critical domains like robotics, aerospace, smart healthcare, and autonomous driving.

The realm of knowledge on this subject is notably broad, despite its relatively recent emergence, as evidenced by recent surveys ([4]–[8]). This panel seeks to bring together the diverse contributors from the scientific community working in this field. The primary goal is to engage them and the audience in discussions about the significant achievements to date and chart the course for future developments, presenting them as open challenges to be tackled.

Acknowledgment

This work has been co-founded by the ANR France 2030 AdaptING project, "ANR-23-PEIA-0009".

References

[1] Y. LeCun, Y. Bengio, G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436-444, May 2015.
[2] B. Moons, et al, "14.5 Envision: A 0.26-to-10TOPS/W subword-parallel dynamic-voltage-accuracy-frequency-scalable Convolutional Neural Network processor in 28nm FDSOI," in Proc. Int. Solid-State Circuits Conf., pp. 246-247, 2017.
[3] European Commission. Directorate General for Communications Networks, Content and Technology., The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self assessment. LU: Publications Office, 2020.
[4] C. Torres-Huitzil and B. Girau, "Fault and Error Tolerance in Neural Networks: A Review," IEEE Access, vol. 5, pp. 17322-17341, 2017.
[5] S. Mittal, "A survey on modeling and improving reliability of DNN algorithms and accelerators," Journal of Systems Architecture, vol. 104, p. 101689, 2020.
[6] A. Ruospo, E. Sánchez, L. Matana Luza, L. Dilillo, M. Traiola, A. Bosio, "A Survey on Deep Learning Resilience Assessment Methodologies," Computer vol. 56 no. 2, pp. 57-66, 2023.
[7] M. Hasan Ahmadilivani, M. Taheri, J. Raik, M. Daneshtalab, M. Jenihhin, "A Systematic Literature Review on Hardware Reliability Assessment Methods for Deep Neural Networks", arXiv:2305.05750, 2023, in review.
[8] C. Bolchini, L. Cassano, A. Miele, "Resilience of Deep Learning applications: a systematic survey of analysis and hardening techniques", arXiv:2309.16733, 2023, in review.