Contribution ID: **10**

Type: **not specified**

Evaluating the Reliability of AI Accelerators with Proton Experiments

Applications based on Artificial Intelligence (AI) are being employed in several fields, including medical diagnosis, drug discovery, robotics, and autonomous vehicles. Enabling robots and vehicles to perform tasks autonomously can be useful in several scenarios, particularly for space exploration and satellites [1]. This success can be attributed in great part to advances in AI models, such as Convolutional Neural Networks (CNNs) and, more recently, Transformer models [2,3].

Due to their size and complexity, modern AI models require parallel accelerators, such as Graphics Processing Units (GPUs) or Tensor Processing Units (TPUs). The architectures of these devices are optimized to perform highly parallel computations efficiently by employing a large number of computing cores. While this kind of architecture can accelerate parallel applications, a fault in a shared or critical resource may affect multiple output elements. Additionally, the large number of computational resources in accelerators further increases the likelihood of SEUs. Unfortunately, radiation-hardened devices are extremely expensive, offer poor performance, and have low power-efficiency, which make them unsuitable to run large AI models. Therefore, in order to safely deploy Commercial-Off-The-Shelf (COTS) devices in space, it is mandatory to ensure both reliability and availability.

Our past research has evaluated the reliability of several AI accelerators under different types of radiation [4,5]. This includes measuring the response of devices to radiation-induced Single Event Latchups (SELs), the effects of Total Ionizing Dose (TID), and creating realistic fault models based on Single Event Upsets (SEUs) observed during experiments with radiation. Further, our previous work has analyzed how faults propagate through AI models, providing insights into the effects of radiation on AI applications [5]. Despite continuous efforts into this field of research, there are several aspects of radiation-induced effects that must be more thoroughly evaluated.

The Trento Proton Therapy Center facility could provide interesting experimental data, advancing the knowledge of the reliability of AI accelerators against protons. Performing experiments at the facility would allow us to evaluate state-of-the-art AI models with new and efficient AI accelerators in realistic settings. For instance, as shown in Figure 1, in previous beam experiments we investigated how Silent Data Corruption (SDC) propagates through the layers of AI models, such as the Data-efficient image Transformers (DeiT). These results showed that radiation-induced SDCs start as errors with relatively high magnitude (red bars), which are then spread to subsequent layers. As the error spreads, however, the error magnitude quickly reduces (red bars), while the ratio of affected elements (blue bar) quickly increases. In other words, AI models quickly dilute the error into the elements of a layer, taking few layers to propagate the error to most elements. With further experiments, we could better identify the root causes of radiation-induced misclassifications in AI models.

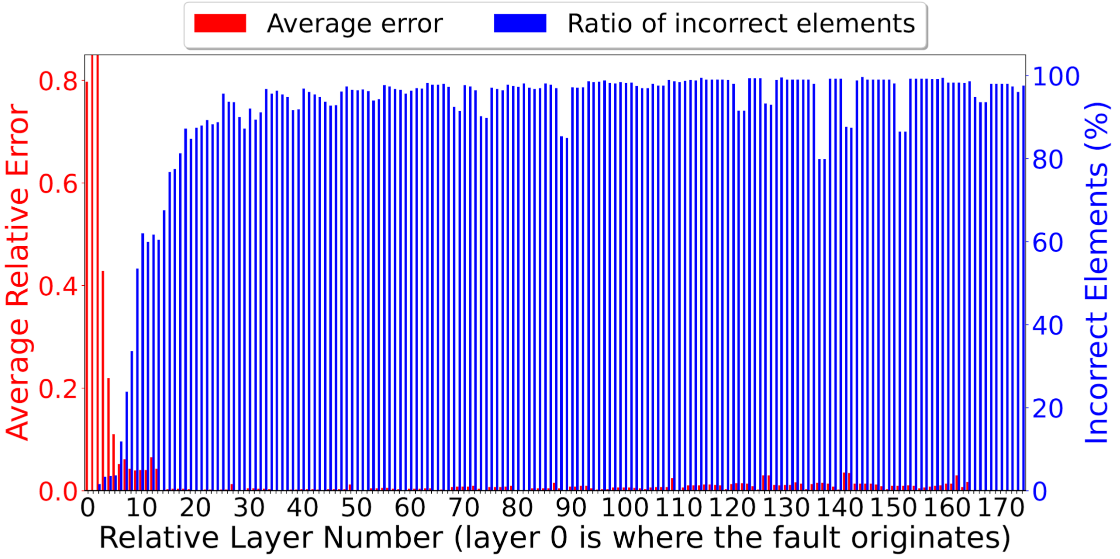

Figure 1: Propagation of SDCs throughout the layers of Data-efficient image Transformers (DeiT).

Bibliography
[1] P. Rech et al., “Impact of GPUs parallelism management on safety-critical and HPC applications reliability”
, in IEEE/IFIP DSN, 2014.
[2] K. He et al., “Deep residual learning for image recognition”, 2015.
[3] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,”2021.
[4] F. F. dos Santos et al., “Experimental findings on the sources of detected unrecoverable errors in gpus”, IEEE Transactions on Nuclear Science, 2022.
[5] P. R. Bodmann et al., “Neutrons sensitivity of deep reinforcement learning policies on edgeai accelerators,” IEEE Transactions on Nuclear Science, 2024.

Primary authors: LOUREIRO COELHO, Bruno (University of Trento); Dr RECH, Paolo (University of Trento)

Presenter: LOUREIRO COELHO, Bruno (University of Trento)