Neutron-Induced Error Rate of Vision Transformer Models on GPUs

Fernando Fernandes dos Santos, Paolo Rech, Angeliki Kritikakou, Olivier Sentieys

Neutron-Induced Error Rate of Vision Transformer Models on GPUs

Fernando Fernandes dos Santos*, Paolo Rech+, Angeliki Kritikakou*, and Olivier Sentieys*
*Univ. Rennes, INRIA, France
+University of Trento, Italy

***Abstract*—Vision Transformers (ViTs) are the new trend to improve performance and accuracy of machine learning. Through neutron beam experiments we show that ViTs have a higher FIT rate than traditional models but similar error criticality.**

I. Introduction

In recent years, the artificial intelligence field has witnessed a surge in very large machine learning models. Among them, the Transformer model has emerged as a powerful tool that outperforms traditional Convolutional Neural Networks (CNNs) in various tasks, including natural language processing, image classification, and instance segmentation [1], [2]. Unlike CNNs, transformers can effectively extract the essential information in a given frame rather than treating all pixels equally. Moreover, Transformers are highly scalable and can be parallelized more efficiently, making them an attractive choice for large-scale image processing tasks. Transformers can also learn a wide range of concepts from the data, which is particularly useful when dealing with complex image datasets.

In order to train and use large Transformer models, GPUs are currently the most suitable hardware architecture. GPUs have evolved from being dedicated to gaming, graphics, and video rendering, to being flexible accelerators for various High-Performance Computing (HPC) and safety-critical applications, including Machine Learning (ML). To meet the demands of these different markets, GPU vendors have made significant advances in terms of computing capabilities and efficiency in programming frameworks. However, reliability becomes a critical concern when deploying GPUs to execute Vision Transformers (ViT) in safety-critical domains. To take advantage of the benefits of ViT in safety-critical systems, ViT-based systems must meet the strict requirements of the ISO26262 and ISO/PAS 21448 standards. In this context, radiation-induced soft errors pose a particularly significant threat as they have been found to have high error rates in commercial devices [3]. Additionally, GPUs have a high fault rate due to their large number of hardware resources [4], [5]. The possibility of having multiple computing units corrupted can significantly undermine the reliability of ViT.

Previous studies have evaluated the reliability of GPUs [6], [7], FPGAs [8], and ASIC accelerators [9] while running various common CNNs. However, none of these studies have comprehensively assessed the reliability of Vision Transformers (ViT) for image classification using neutron beams. This study is the first to assess ViT models' reliability and compare it with commonly used CNNs, ResNet.

The paper is structured as follows: Section 2 provides background on ViT and radiation effects on computing devices. Section 3 details the experimental methodology. Section 4 presents the results, and Section 5 concludes the paper.

II. Background

The Transformers architecture differs significantly from typical CNN models used for image-related tasks. Figure 1 depicts the original Vision Transformer architecture proposed by Dosovitskiy *et al.* in [1]. Unlike traditional CNN models, Transformers input images are shaped as sequences of *patches*. These patches are flattened into a single vector and linearly projected into a dimension, enabling the patches to be fed to the Transformer Encoder. Along with the patches, extra learnable parameters (i.e., the class and position encodings) are embedded to allow the model to learn the image structure.

The Transformer Encoder consists of a block of Layer Normalizations, Multi-Layer Perceptrons (MLP), and Multi-Head Attention (MHA) networks, with the latter being the key innovation of the Transformers models [10]. While the Layer Normalization and MLP blocks are standard algorithms in state-of-the-art deep learning models, the MHA modules enable the Transformer encoder to give *attention* to critical regions of the image, generating attention maps for the received embedded patches. Consequently, attention maps can provide information about the relevant parts of the image and give context to the information. After the encoder, the classification head applies a fully connected layer followed by a softmax to the output, generating the classification labels (i.e., the probabilities of the image's classification).

One of the critical concepts of Transformer models is scalability, which allows them to be trained on large amounts of data and achieve high accuracy. According to the Transformers' original studies [1], [10], Transformers have lower accuracy than ResNet CNNs, when trained on small datasets or not fine-tuned with diverse datasets. In order to achieve maximum accuracy, Transformers must be trained on large datasets and then fine-tuned with a different dataset.

Since Transformers can push GPUs to their limits, it is reasonable to expect that ViT models will have a higher error rate than common CNNs. This is because the fault probability and resource usage are linearly dependent. Moreover, it is worth noting that the number of patches and input size affects the accuracy and performance of the ViT model. Therefore, it is also likely that those parameters may affect the model's reliability. Smaller patches might grant the model higher reliability, as an error in a small patch, that does not have essential

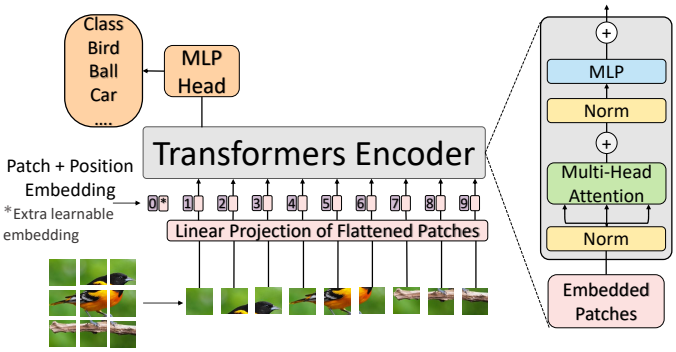

Fig. 1: The architecture of the Transformer model used in the experiments. The input image is divided into a sequence of linear patches and passed through the Encoder, which consists of a series of operations, including normalization, MLP, and the attention mechanism. The figure was extracted from [1].

information, will not be used for the final classification. This behavior has been observed in CNN in previous works [6]. Whereas a highly complex CNN, such as Faster R-CNN, can better deal with errors than a less complex CNN, like YOLO. In Section IV, we discuss these aspects.

GPUs executing Vision Transformers are susceptible to errors caused by neutrons. Terrestrial neutrons can interact with electronic devices, leading to bit-flips in memory or current spikes in logic circuits, which can generate soft errors [3]. While soft errors typically do not harm the device, they can impact the output of an ML model and alter its final classification. On GPUs, as with any other device, faults propagate from the hardware to the software level, potentially leading to the following outcomes:

1) **Masked**: The corrupted data is not used, or the circuit functionality is not affected, so there is no effect on the program output.
2) **Detected Unrecoverable Error (DUE)**: The program stops working, or the entire system crashes, which causes a disruption to the program execution.
3) **Silent Data Corruption (SDC)**: The application appears to finish successfully, but the output is incorrect, and no flag or indication of an error is set, which can go unnoticed if the output is not properly validated. The SDCs on ViTs models, and any ML model, can be divided into two classes: (1) **Tolerable SDCs** that change the output of the Transformers MLP head but not the classification, and (2) **Critical SDCs** that modify the classification entirely (i.e., the top 1 classification probability is changed). Critical SDCs are especially hazardous for safety-critical applications, as an incorrect inference can have catastrophic consequences.

III. Evaluation Methodology

In this section, we describe our experimental methodology for evaluating the ViT models, as well as the metrics we use to assess the error rates. We also provide detailed information on each code characteristic and explain how such aspects could potentially impact fault propagation.

TABLE I: Evaluated model details. The accuracy for the full dataset is extracted from the original paper.

	Model size [MB]	Execution time [s]	Accuracy
ResNet 50	97.70	0.56	80.86%
ViT 16/224	330.23	5.30	84.53%
ViT 32/224*	336.55	0.84	80.30%
ViT 32/384	336.83	3.88	81.66%

A. Codes and device under test

For our evaluations, we select two types of ML models: ResNet50 and Vision Transformers (ViT). We used the Imagenet dataset [11] as input, which comprises 14 million images divided into 1,000 classes. We chose a subset of 256 images of the Imagenet dataset to perform the experiments efficiently. Then, we group the images into four batches of 64 images and fed each batch to the model with the goal of stressing all the GPUs' functional units. For all evaluations, we use PyTorch 1.13.0 with CUDA 11.7.

For the evaluation of the ViT models, we chose two configurations from the baseline ViT models proposed in the original paper [1] (ViT 16/224 and ViT 32/384) and one additional configuration with SAM optimizations (ViT 32/224*). The names of the models identify their characteristics, which include the size of the patches and the size of the input frame. For instance, ViT 16/224 splits an image of size 224×224 into patches of size 16×16. It is worth noting that the model becomes more accurate as the patch size gets smaller and the input size increases. All the ViT are pre-trained models from the Timm/HuggingFace model database [12]. This database consists of an open-source collection of many transformer models, including popular ones, such as BERT and GPT4.

To provide a comprehensive evaluation, we compare the reliability of ViT with a conventional CNN model, ResNet 50. ResNet is a Residual Neural Network that combines convolutional layers with residual units. ResNet is one of the most widely used CNNs for classification, as it can achieve satisfactory accuracy even with quantized models. In this work, we use a ResNet composed of 50 layers with a nominal accuracy of 80.86% on the Imagenet dataset. The ResNet 50 model is a pre-trained model embedded with PyTorch vision models. Table I shows the details of the ML models we evaluate, i.e., Model size in MB (i.e., the amount of memory that the model allocates on the GPU), Execution time in seconds (for a batch of 64 images), and Accuracy from the original papers. In fact, the ResNet 50 is 3.4× smaller than the smallest ViT model. One of the main characteristics of Transformers is being large by default, meaning that in order to increase accuracy, they have to increase the number of parameters of the models.

We conducted an evaluation of the reliability of ML models running on NVIDIA Maxwell GPUs, specifically the Quadro M2000 model. This GPU is based on the Maxwell Instruction Set Architecture (ISA) and is fabricated with TSMC CMOS technology using a $28nm$ process node with 4GB of DDR memory, 768 CUDA cores, and power consumption of 75W at a frequency of operation of 796 MHz.

B. Neutrons Beam Experiments

To measure the FIT rate, we take advantage of controlled neutron beam experiments. Beam experiments were performed at the ChipIR facility of the Rutherford Appleton Laboratory (RAL), UK. The available neutron flux at ChipIR is about $10^6 n/(cm^2/s)$, i.e., about 7-8 orders of magnitude higher than the terrestrial flux ($13n/(cm^2 \times h)$ at sea level [13]). Since the terrestrial neutron flux is low, in a realistic application, it is highly unlikely to observe more than a single corruption during the program execution. We have carefully designed the experiments to maintain this property (observed error rates were lower than (1 error per 700 executions). Experimental data, then, can be scaled to the natural radioactive environment without introducing artifacts.

Figure 2 shows the setup installed at ChipIR. We align the GPUs with the neutron beam and control them using motherboards. To protect the supporting motherboards and equipment from scattering thermal neutrons, they are covered with boron plastic. Only the GPU core is exposed to radiation, and the board's DDR and power control circuitry are outside the beam spot. The purpose is to study the radiation effects on the computing core of the GPU, not on the DDR, which has already been studied extensively. The software setup comprises Python scripts that run on a server computer outside the beam room. The watchdog monitors, executes the ML models, logs events, and recovers from device hangs. If the program stops responding within a predefined interval (up to 10× the expected execution time), it is killed and relaunched. The software setup executes the same ML model inside the GPU for a predetermined number of iterations. After each iteration, the output is compared with a constant golden value (the expected values for the last layer of the models). If there is a mismatch between the outputs, the server logs the relevant information and restarts the model executions. The setup logs whether the errors significantly alter the classification probabilities and generate a misclassification (i.e., a Critical SDC). SDCs that only modify the output of the last model's layer but maintain correct classification are categorized as Tolerable SDCs.

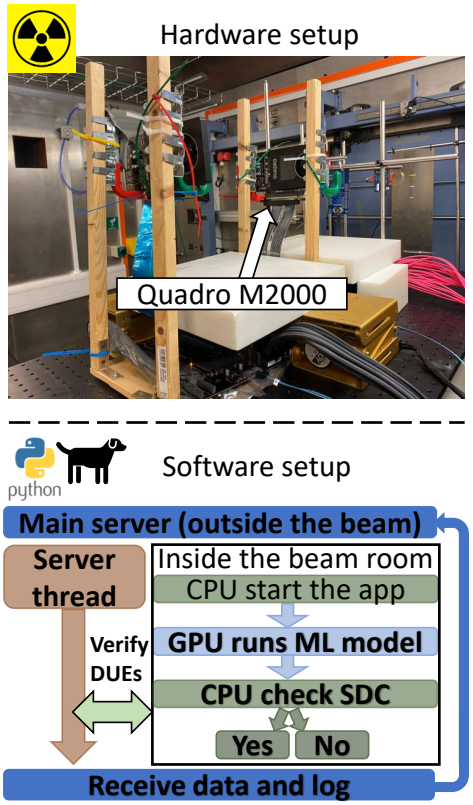

Fig. 2: The experimental hardware setup comprises GPUs aligned along the ChipIR beamline, and motherboards control the GPUs. The software setup involves a Python-based server that manages GPUs within the beam room.

IV. Experimental Results

This section presents the first-ever measured error rate of Transformers models under neutron beam irradiation on GPUs. We also compare the error rate of the ViT model with the commonly used ResNet 50 DNN. Furthermore, we analyze which types of events significantly impact the DUE rate of the evaluated models.

Figure 3 shows the Tolerable Silent Data Corruption (SDC), Critical SDC, and Detected Unrecoverable Error (DUE) FIT rates (y-axis) for the four configurations evaluated (x-axis). The error rates are presented with a Poisson distribution with a 95% confidence interval.

As expected, the SDC FIT rates for the ViT configurations are similar to or much higher than the ResNet 50 rate. Specifically, the SDC rates are 71.60 for ResNet 50, 101.81 for ViT 16/224, 372.36 for ViT 32/224, and 398.97 for ViT 32/384. Due to the high GPU resource demands of the ViT models, the functional units and thread schedulers are heavily stressed, leading to higher error rates. Additionally, as the ViTs models put high pressure on the GPUs memories, and the Quadro M2000 GPUs do not have Error Correction Codes to protect the register file, shared, and cache memories, it is then expected that the error rate increases with the memory usage.

One exciting result from Figure 3 is that the error rate of ViT 16/224 is significantly lower than the other two configurations (3.66× lower than 32/224 and 3.92× lower than 32/384). We attribute this difference to the model architecture, which has the smallest patch size among the tested ViTs. With more patches per frame, the probability of faults disturbing the MLP head output decreases, which may mask some of the SDCs.

All tested ML models have a Critical SDC rate much lower than the Tolerable SDC rate. For ResNet 50, ViT 16/224, ViT 32/224, and ViT 32/384, Critical SDC rates are 15.91, 10.43, 16.93, and 14.91, respectively. Notably, the ViT model with the smallest patch size has the lowest Critical SDC rate across all tested configurations. This can be explained by the fact that a model that splits the image into smaller patches may be less affected by a fault in one of the attention modules that compute the attention maps for a given patch. In the final paper, we will investigate whether this trend is consistent across other ViT models with different patch sizes and encoder configurations, such as ViT-G/14, SwinV2-G, and DaViT-G.

The DUE rates, on the other hand, are similar among the tested models (15.91 for ResNet 50, 10.43 for ViT 16/224,

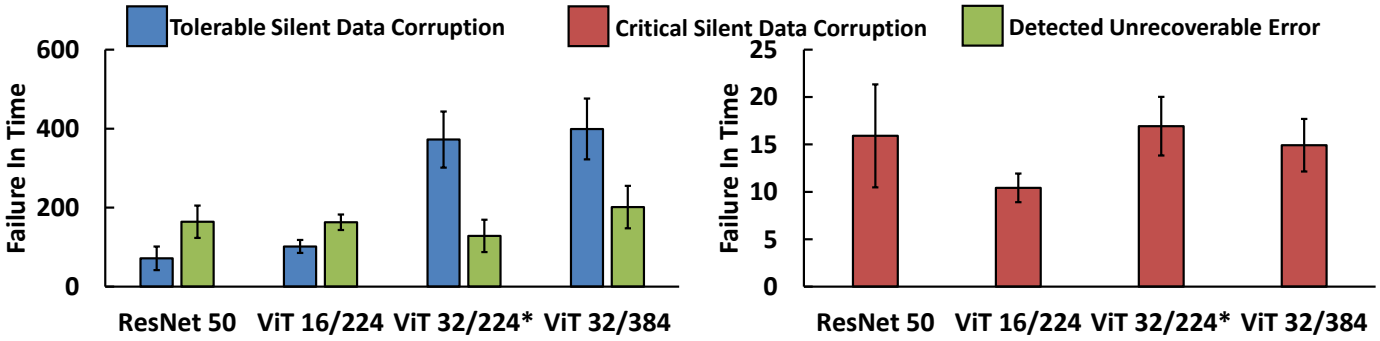

Fig. 3: Tolerable and Critical SDC and DUE FIT rates for ResNet 50 and 3 ViT configurations (16/224, 32/224, and 32/384).

TABLE II: DUE sources for all tested configurations.

	ResNet 50	ViT: 16/224	32/224*	32/384
Illegal Instruction	4.8%	1.9%	12.8%	1.8%
Misaligned/illegal memory address	45.2%	36.8%	59.0%	28.6%
Launch Errors	6.5%	15.8%	10.3%	5.4%
System Crash/Hang	43.5%	45.5%	17.9%	64.3%

16.93 for ViT 32/224, and 14.91 for ViT 32/384). The DUE rate is less dependent on the algorithm's computational complexity and more on events such as illegal instructions, system crashes/hangs (i.e., GPU hangs or the driver of the GPU makes the operating system not usable), incorrect addresses for jump and branch instructions, kernel launch errors, synchronization between CPU-GPU, and illegal memory access. Table II presents the percentage of different sources that lead to DUEs. Not surprisingly, System Crash/Hang and Misaligned/Illegal memory accesses are the primary sources of DUEs, accounting for an average of 85.2% of the DUEs. An error on a register storing a memory address will likely generate a crash or make the program enter an infinite loop.

The communication between CPU-GPU is also a significant source of system crashes and directly depends on the number of kernels launched by the ML framework. For the evaluated models, PyTorch uses basic operations (Torch NN modules) to define a graph representing the ML model. Although the models may differ in complexity, the number and types of basic modules are similar, with different tensor sizes and parameters. ResNet 50 has 158 modules, while all the ViT models have 187. Each module requires at least one device-host (CPU-GPU) synchronization, and a transient fault during these synchronizations could potentially result in a GPU DUE.

V. Conclusions

In this study, we have analyzed the error rate and criticality of three ViT models and compared them with a ResNet CNN. While ViT is currently the state-of-the-art ML model for vision tasks, they are highly resource-demanding and have shown a high error rate. Further investigation is required to understand how the characteristics of the Transformer model, such as patch size and encoder structure, impact the error criticality. As the use of ViT increases, understanding the factors contributing to their error rate and criticality will be crucial in developing effective strategies to mitigate critical errors.

Acknowledgment

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the grant agreements No 899546 (Marie Skłodowska-Curie) with the support of the Brittany Region and 101008126 (RADNEXT). It was partially funded by ANR FASY (ANR-21-CE25-0008-01) and ANR RE-TRUSTING (ANR-21-CE24-0015-02). ChipIR provided and supported neutron beam time experiments (DOI 10.5286/ISIS.E.RB2200303). We acknowledge Dr. Christopher Frost, Dr. Maria Kastriotou, and Dr. Carlo Cazzaniga for their help with beam experiments.

References

[1] A. Dosovitskiy *et al.*, *CoRR*, 2020. arXiv: 2010.11929.
[2] S. Khan *et al.*, *ACM Comput. Surv.*, 2022.
[3] R. Baumann, *IEEE Design Test of Computers*, 2005.
[4] D. A. G. Oliveira *et al.*, *IEEE Trans Comput*, 2016.
[5] M. B. Sullivan *et al.*, in *IEEE/ACM International Symposium on Microarchitecture*, 2021.
[6] F. F. d. Santos *et al.*, *IEEE Trans. Reliab.*, 2019.
[7] S. K. S. Hari *et al.*, *IEEE Transactions on Dependable and Secure Computing*, 2021.
[8] F. Libano *et al.*, *IEEE Trans. Nucl. Sci.*, 2019.
[9] R. L. Rech Junior *et al.*, *IEEE Trans. Nucl. Sci.*, 2022.
[10] A. Vaswani *et al.*, *CoRR*, 2017. arXiv: 1706.03762.
[11] J. Deng *et al.*, in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009.
[12] R. Wightman, https://github.com/huggingface/pytorch-image-models.
[13] JEDEC, JEDEC Standard, Tech. Rep., 2006.