International Journal of Electrical and Computer Engineering (IJECE)
Vol. 11, No. 5, October 2021, pp. 4481~4488
ISSN: 2088-8708, DOI: 10.11591/ijece.v11i5.pp4481-4488

Analyzing the instructions vulnerability of dense convolutional network on GPUS

Khalid Adam[1], Izzeldin I. Mohd[2], Younis Ibrahim[3]
[1,2]College of Engineering, University Malaysia Pahang, Malaysia
[3]College of IoT Engineering, Hohai University, China

Article Info

Article history:

Received Sep 4, 2020
Revised Mar 10, 2021
Accepted Mar 21, 2021

Keywords:

DenseNet201
GPUs
Healthcare
Reliability
Soft error

ABSTRACT

Recently, deep neural networks (DNNs) have been increasingly deployed in various healthcare applications, which are considered safety-critical applications. Thus, the reliability of these DNN models should be remarkably high, because even a small error in healthcare applications can lead to injury or death. Due to the high computations of the DNN models, DNNs are often executed on the graphics processing units (GPUs). However, the GPUs have been reportedly impacted by soft errors, which are extremely serious issues in the healthcare applications. In this paper, we show how the fault injection can provide a deeper understanding of DenseNet201 model instructions vulnerability on the GPU. Then, we analyze vulnerable instructions of the DenseNet201 on the GPU. Our results show that the most significant vulnerable instructions against soft errors PR, STORE, FADD, FFMA, SETP and LD can be reduced from 4.42% to 0.14% of injected faults, after we applied our mitigation strategy.

Corresponding Author:

Khalid Adam
Faculty of Electrical and Electronic Engineering Technology
University Malaysia Pahang
26300, Pahang, Malaysia
Email: khalidwsn15@gmail.com

1. INTRODUCTION

Recently, the success of deep neural network (DNN) in challenging perception tasks makes them a powerful tool for many applications, including safety critical system such as healthcare application [1]. For instance, DNN is used in surgical procedures where it gives a better understanding of surgical practices and automatic recognition of workflow [2], [3]. This and other potential DNN applications are associated with high risk of humam injury or death, especially if there is a malfunction. Hence, it should be exceptionally reliable [4]-[6]. However, with the growing complexity of modern digital hardware platforms (i.e., GPUs), it has become increasingly difficult to guarantee the reliability of hardware operations when DNN models are run on top of the hardware (i.e., GPUs) [7]-[9]. Notably, soft errors are caused by a transient signal. This is induced by a single energetic particle strike when the collected charge is greater than the critical charge required to cause a change in the state of a memory cell, register, latch, or flip-flops [10], [11]. As a result, this could eventually lead to misclassification of objects in DNNs, and the consequences would be disastrous. Therefore, when the DNNs is performed in GPUs, their reliability implication is not well understood in the healthcare applications, because the errors propagate from the GPUs to DNNs (i.e., DenseNet201) [7], [12].

Several techniques have been proposed to reduce the soft errors in the GPUs such as double modular redundancy (DMR), triple modular redundancy (TMR), and algorithm-based fault tolerance (ABFT) [13], [14]. Nevertheless, the main issue with these solutions is that they have runtime overhead and are not cost