Interactive Presentations (IP)
Topic 7 - Interactive Presentations (7)

Author: Mr. Shen Li
Xi'an Microelectronics Technology Institute, China Aerospace Science and Technology Corporation (CASC), China, lishen565@126.com

Dr. Zhong Ma
Xi'an Microelectronics Technology Institute, China Aerospace Science and Technology Corporation (CASC), China, mazhong@mail.com
Mr. Zhuping Wang
Xi'an Microelectronics Technology Institute, CASC, China, zxjwl@126.com
Prof. Xuehan Tang
Xi'an Microelectronics Technology Institute, China Academy of Space Electronics Technology (CASET), China Aerospace Science and Technology Corporation (CASC), China, m13772060157@163.com

A SOFTWARE FAULT-TOLERANT METHOD FOR SATELLITE-BORNE NEURAL NETWORK ACCELERATOR

Abstract

Deep Neural Networks (DNN) has proved its potential in various perception tasks and becomes an appealing option for space missions with increasing complexity. But the DNNs require a large number of arithmetic operations, the satellite intelligent applications aren't available without specific neural network accelerator. However the satellite-borne neural network accelerator is very difficult to develop, because in order to tackle the computational error caused by the Single Event Upset (SEU), using traditional hardware-based fault-tolerant methods, such as Triple Modular Redundancy (TMR), are costly and have a large volume and power consumption, and it will greatly reduce the computing performance of the neural network accelerator so that make it difficult to meet the actual application requirements. Therefore, it is necessary to design a software fault-tolerant method to minimize the influence of the traditional fault-tolerant method upon the performance of the neural network accelerator. In this paper, we proposed a software fault tolerant scheme for the neural network accelerator, which effectively overcomes the shortcomings of the hardware-based methods. This method comprises of two major aspects. Firstly, during the training process of a neural network model, some bit noise is added to weights and feature maps, which in order to simulate the situation of SEU in the space environment, so as to further enhance the intrinsic robustness of the neural network model. Secondly, during the inferencing process of a neural network on the satellite-borne neural network accelerator, the inference data is automatically expanded to two parallel redundant computing flows for the same input data. At the appropriate layers, such as the activation layer, the error-checking operation by element-wise comparison is adapted to determine which data has been flipped caused by SEU. If a difference between a pair of two elements from two parallel flow is detected, the element with smaller deviation from its surrounding area is considered as the correct element under the smoothness assumption of data in the neural network algorithm. The proposed fault-tolerant method is a software method which can effectively improve the fault-tolerant capability of satellite-borne neural network accelerator, greatly reduce the impact of the traditional fault-tolerant method on the performance, and pave the way for using various commercial neural network accelerators without radiation protection on the satellite.