Understanding Error Propagation in Deep

Learning Neural Network (DNN) Accelerators and Applications-2017

Based on our observations, we discuss the reliability implications of designing DNN accelerators and applications and also propose two cost-effective error protection techniques to mitigate Silent Data Corruptions (SDCs) i.e., incorrect outcomes. The first technique, symptom-based detectors, is implemented in software and the second technique, selective latch hardening, in hardware. We evaluate these techniques with respect to their fault coverage and overhead.

**DNN Accelerator**

Many specialized accelerators [12, 13, 25] have been proposed for DNN inferencing, each with different features to cater to DNN algorithms. However, there are two properties common to all DNN algorithms that are used in the design of all DNN accelerators: (1) MAC operations in each feature map have very sparse dependencies, which can be computed in parallel, and (2) there are strong temporal and spatial localities in data within and across each feature map, which allow the data to be strategically cached and reused.

**Silent Data Corruption (SDC)**

We define the SDC probability as the probability of an SDC given that the fault affects an architecturally visible state of the program (i.e., the fault was activated). This is in line with the definition used in other work [20, 26, 36, 59]. In a typical program, an SDC would be a failure outcome in which the application’s output deviates from the correct (golden) output. This comparison is typically made on a bit-by-bit basis. However, for DNNs, there is often not a single correct output, but a list of ranked outputs each with a confidence score as described in Section 2.1, and hence a bit-by-bit comparison would be misleading. Consequently, we need to define new criteria to determine what constitutes an SDC for a DNN application.We define four kinds of SDCs as follows:

• SDC-1: The top ranked element predicted by the DNN is different from that predicted by its fault-free execution. This is the most critical SDC because the top-ranked element is what is typically used for downstream processing.

• SDC-5: The top ranked element is not one of the top five predicted elements of the fault-free execution of the DNN.

• SDC-10%: The confidence score of the top ranked element varies by more than +/-10% of its fault-free execution.

• SDC-20%: The confidence score of the top ranked element varies by more than +/-20% of its fault-free execution.