ABSTRACT

In the era of artificial intelligence, deep neural networks, image and video processing applications are very power consuming and require lot of memory to store the parameters and metadata. Approximate compute and approximate memory units are promising solutions which provides benefits in terms of power, area and memory requirement. Approximate units exploit the error resilient behaviour of these applications, which result in significant benefits in power and area at a minimal cost of slightly degraded output.

In this research work, we analyse four biased restoring array dividers (BRAD) targeted towards error resilient applications. The basic repeating unit in BRAD is a controlled subtractor unit (CSU). Four different biased controlled approximate subtractor units(BCSUs) that have a maximum of only one error in the borrow bits of the subtractor are analysed. The CSU is replaced in dividers with BCSUs to obtain BRADs using three different replacement strategies – horizontal, vertical, and triangular. We analyzed 96 BRAD designs and compared them with 192 existing approximate divider designs for three image processing applications. We obtain the pareto-optimal designs for the same with respect to various performance metrics like power, area, delay, and power delay product. We show that if we limit the maximum reduction in SSIM to 20%, one of the proposed biased divider design BRAD3 outperforms all the existing designs in the literature considered for this experiment.

This work also compares Quantised Non-Standard float point(qNSFP) representation with Posit and Fixed-Posit number formats for a wide variety of pre-trained Deep Neural Networks (DNNs). We observe that fixed posit representation is far more suitable for DNNs as it results in faster and low-power computation circuit. We show that

posit quantisation achieves accuracy within the range of 0.3% of Top-1 accuracy with a significant benefit in terms of memory compression, almost 20%, as compared to state of the art qNSFP, and achieves better accuracy in almost all models compared with qNSFP technique for the same bit-width. For fixed posits, on an average, accuracy remains within the range of 0.57% of Top-1 accuracy. We have implemented posit and fixed-posit based multipliers using 28nm FDSOI technology node. We observe that posit based multiplier require more PDP and area compared to qNSFP for the same bit-width. However, for fixed-posit we observe upto 71% reduction in PDP and upto 36% reduction in area for the same bit-width. We can achieve similar PDP and area by decreasing the posit bit-width compared to qNSFP with a gradual drop in Top-1 accuracy, unlike the qNSFP where the accuracy drops drastically. We further show that larger DNN can be quantised more with very little accuracy drop.

We also analyse the Mixed-V_T 8T SRAM architecture for 5 different neural networks. Our proposed Mixed-V_T 8T SRAM architecture requires maximum of 0.17x(0.46x) and 0.28x(0.69x) dynamic energy(leakage power) than Het-6T and Hyb-8T/6T SRAM architecture respectively at 0.5V and maximum of 0.35x(0.84x) and 0.46x(0.90x) dynamic energy(leakage power) than Het-6T and Hyb-8T/6T SRAM array respectively at 0.7 V for 6-bit weights of neural networks.