RELIABLE CROSS-LAYER APPROXIMATIONS IN DEEP LEARNING

Approximate computing has shown significant efficiency gains with respect to power, chip area, and performance/ throughput by exploiting the inherent error resilience of applications, such as Deep Learning (DL). It works on the principle of relaxed precision, i.e., trading off tolerable accuracy loss to enable novel optimizations, keeping a good balance between quality, reliability, and energy efficiency. Due to this, researchers have employed approximate computing techniques at various levels of the DL computing stack.

The Data-Level approximation techniques aim to achieve the desired quality-energy metrics by simplifying the processing of data within deep network layers. Alternatively, approximate data representations also aim to reduce the complexity and volume of datasets. The data-level techniques include sampling, quantization, and relaxed precision (such as post-training quantization and quantization aware training), and compression, etc. Whereas the software-level techniques focus on the algorithmic part of the model, primarily by skipping unnecessary computations based on certain pre-determined criteria. Examples of some of the software-level techniques involve pruning, iterative refinement, loop perforation, sparsity, and knowledge distillation etc. Similarly, at the hardware level, approximate adders, multipliers, and MACs have been employed in the compute units for DL processing.

While researchers have employed these techniques in isolation (at the distinct levels), cross-layer approaches (employing the above-mentioned techniques simultaneously to the same DL architecture) need further exploration. Moreover, the wide applicability of approximate DL in the energy-efficient implementation on resource-constrained devices in mission-critical systems requires reliability analysis/ mitigation as part of the cross-layer frameworks.

This research includes the following steps:

1. Review the concerned literature to identify the state-of-the-art cross-layer approaches, along with reliability analysis and mitigation techniques.

2. Select suitable DL models and Datasets.

3. Employ approximations at various levels of the DL computing stack. Identify an optimal combination of approximation at all levels.

3. Proof of concept implementation of the approximated model on FPGA/ ASIC/ other resource-constrained hardware platforms to evaluate the gains in area, power/ energy, and performance, etc.

4. Reliability analysis/ mitigation of the proposed approximations.

Contact CAES Research Group: Yawar Rasheed (y.rasheed@utwente.nl), Ghayoor Gillani (s.ghayoor.gillani@utwente.nl), and Marco Ottavi, m.ottavi@utwente.nl

Note: The theme of this MSc. thesis assignment is flexible and can be adapted to the specific interests of the students.