DESIGN SPACE EXPLORATION OF DEEP LEARNING ALGORITHMS FOR ENERGY EFFICIENT APPROXIMATE ARCHITECTURES

One of the significant challenges that researchers of approximate computing face is estimating the error resilience of the applications and quantifying the effects of errors on the quality and reliability aspects. It is essential to know how much error a specific application can tolerate, so that an appropriate level of approximation can be induced. Various researchers relied on the basic idea of the intrinsic error resilience of Deep Learning (DL) architectures without actually quantifying it. In contrast, certain researchers have employed NSGA-II, White Gaussian Noise, and simulation-based Error resilience analysis methodologies to quantify the resilience of DL algorithms before applying approximations.

Employing approximate computing to DL poses an additional challenge, which is the huge design space of: (1) Segments of the design/ algorithm that are potential candidates for approximations, (2) Approximate arithmetic units that need to be searched, navigated, and evaluated for quality/ energy-efficiency balance. In simple words, this huge design space arises from the fact that a multitude of computations, arithmetic components, and processes may be approximated, each having different implications on the quality, reliability, and energy efficiency. A comprehensive error resilience analysis necessitates exploring and navigating this huge design space efficiently to ensure that approximations are within tolerable limits. Researchers have employed evolutionary algorithms, ML-based search algorithms, and custom algorithms to navigate the design space of DL architectures.

Keeping in view the wide adoption of DL architectures and the increased interest in approximate implementations on resource-constrained devices, this domain requires further exploration of efficient DSE methodologies/ frameworks for effective error resilience analysis. Moreover, it's interesting to quantify the effects of errors on the reliability, in addition to the quality and energy efficiency of the DL architectures.

This assignment includes the following Steps:

1. Understanding the error resilience (and its relation to quality and reliability), the design space of DL architectures, and how to navigate/ evaluate the design space for effective resilience analysis through a review of the literature.

2. Select suitable DL models along with datasets. Perform the error resilience analysis along with the DSE to identify the segments of the model that are most resilient to errors/ approximations and optimal approximate units that can be replaced with exact units.

3. Perform proof of concept implementation by introducing approximations to the identified segments of DL and measure the effect on quality/ accuracy and reliability.

4. 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.

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.