Error Resilience Analysis Of Transformers Approximations For Computer Vision Tasks

Studies: BSc/MSc. Technical Computer Science

Vision Transformer (ViT) models with a novel self-attention mechanism have been proposed as a viable alternative to Convolutional Neural Networks (CNNs). These models have achieved state-of-the-art performance on various computer vision tasks. However, the inherent computational complexity of these newly proposed architectures, coupled with high power and energy requirements limit their application to resource-constrained embedded/ edge devices.

In recent years, approximate computing has gained particular attention as a practical approach for achieving energy efficiency and reducing computational complexity. It exploits the inherent error tolerance of applications, enabling a significant reduction in power consumption and chip area while improving performance and throughput. By relaxing the precision and trading off tolerable accuracy in error-resilient applications, it offers novel optimizations, keeping a good balance between quality, reliability, and energy efficiency.

To reduce the computational and energy demands of Deep Learning (DL) workloads and to enable their efficient deployment on modern resource-constrained and embedded devices, researchers have employed approximation techniques across various levels of the DL computing stack. Some of the widely explored approximation techniques include quantization, sampling at the data-level, pruning and early-exiting at the algorithmic level, and use of approximate multipliers, at the hardware level.

One of the significant challenges that researchers of approximate computing in DL face is estimating the error resilience of the applications and quantifying the effects of errors on the quality. It is essential to know how much error a specific application can tolerate, so that an appropriate level of approximation can be induced. While approximate computing has been explored in the context of ViTs, the comprehensive error resilience analysis of transformer models remains an open question.

Assignment: As part of this research direction, the student will explore the error tolerance of ViTs. A suitable model will be selected, a pretrained version will be implemented using DL frameworks such as PyTorch. Approximations will be investigated by replacing the accurate multipliers with the approximate ones (from public libraries such as evoapproxlib) using the TranxAxx framework (https://doi.org/10.1109/TCASAI.2025.3565685). The resulting tradeoff in quality and efficiency will be analyzed.

Possible Research Questions:

- To what extent are the individual layers of the ViTs resilient to approximate computing?
- What is the quality-efficiency tradeoff in ViTs under approximate multiplier configurations?

Work Division

- Literature: 20%
- Modelling: 30%
- Coding: 20%
- Evaluation: 10%
- Writing: 20%

Contact: Yawar Rasheed (MSc.), CAES research group, y.rasheed@utwente.nl
Dr. ir. Ghayoor Gillani, CAES research group, s.ghayoor.gillani@utwente.nl

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