SETS-2026 Talk

Title: Adaptive Fault Resilience for Early-Exit DNNs (ATS-2025)

Authors:

Rama Mounika Kodamanchili, Natalia Cherezova, Mahdi Taheri, Maksim Jenihhin

Abstract:

Deep learning models are increasingly being deployed in edge and resource-constrained platforms, where reliability concerns become more pronounced. As hardware continues to scale, components such as memory and compute units become more susceptible to transient and permanent faults, which can alter neural network behaviour in unexpected ways. While state-of-the-art DNNs have been studied extensively under such conditions, far less is known about how adaptive or Dynamic Deep Neural Networks (D2NNs) behave when hardware faults occur. Early-exit DNNs are one such class of architectures, designed to reduce computation with predictions at intermediate layers when confidence is sufficient.

In this paper, we investigate the faulty behaviour of early-exit architectures at the model level, evaluated through controlled bit-level fault injections across a range of Bit-Error Rates (BERs). The analysis examines accuracy degradation, changes in exit selection and MAC-based energy cost under fault conditions. This helps in exposing trade-offs between efficiency and resilience of early-exit DNNs as compared to their static counterparts.

I will present these findings and outline how they support the broader goal of developing dependable D2NNs, while also indicating how such model-level insights can guide design strategies for reliable hardware accelerators.