Contribution ID: **47** Type: **Standard Talk**

FKeras: A Sensitivity Analysis Tool for Edge Neural Networks

Wednesday 27 September 2023 14:30 (15 minutes)

Scientific experiments rely on machine learning at the edge to process extreme volumes of real-time streaming data. Extreme edge computation often requires robustness to faults, e.g., to function correctly in high radiation environments or to reduce the effects of transient errors. As such, the computation must be designed with fault tolerance as a primary objective. FKeras is a tool that assesses the sensitivity of machine learning parameters to faults. FKeras uses a metric based on the Hessian of the neural network loss function to provide a bit-level ranking of neural network parameters with respect to their sensitivity to transient faults. FKeras is a valuable tool for the co-design of robust and fast ML algorithms. It guides and accelerates fault injection campaigns for single and multiple-bit flip error models. It analyzes the resilience of a neural network under single and multiple bit-flip fault models. It helps evaluate the fault tolerance of a network architecture, enabling co-design that considers fault tolerance alongside performance, power, and area. By quickly identifying the sensitive parameters, FKeras can determine how to protect neural network parameters selectively.

Primary author: WENG, Olivia

Co-authors: G, Abarajithan (UC San Diego); MEZA, Andres (UC San Diego); HAWKS, Ben (Fermi National Accelerator Lab); CRUTCHFIELD, Christopher (UC San Diego); CAMPOS, Javier; DUARTE, Javier Mauricio (Univ. of California San Diego (US)); TRAN, Nhan (Fermi National Accelerator Lab. (US)); BOCK, Quinlan (Fermilab); KASTNER, Ryan

Presenter: WENG, Olivia

Session Classification: Contributed Talks

Track Classification: Contributed Talks