Robust and Energy-Efficient Deep Learning Systems

Muhammad Abdullah Hanif¹ (Ph.D. Candidate), Muhammad Shafique² (Advisor)
¹Technische Universität Wien (TU Wien), Vienna, Austria
²Division of Engineering, New York University Abu Dhabi (NYUAD), Abu Dhabi, UAE

Overview of Our Methodology

- **Techniques for Robust and Energy-Efficient Deep Neural Network Inference**
  - **SalvageDNN: Fault-Aware Mapping**
    - Filter 1: $W_i = 0.26$, $W_{i+1} = 0.69$, $W_{i+2} = 0.41$, $W_{i+3} = 0.24$
    - Filter 2: $W_i = 0.45$, $W_{i+1} = 0.45$, $W_{i+2} = 0.45$, $W_{i+3} = 0.45$
  - **HW/SW Approximations**
    - For VGG11 with ImageNet dataset, our SalvageDNN method takes $<2$ ms on a GPU while one epoch of re-training takes $>100$ mins with a GPU.

Selected Publications


Software-level Techniques for Mitigating Reliability Threats

- Fault-Aware Training
- Fault-Aware Mapping
- Range Restriction-Based Soft Error Mitigation

Hardware Modifications for Mitigating Reliability Threats

- Permanent Fault Mitigation Support
- Aging Mitigation

Selected Datasets

- Cifar10
- Cifar100
- ImageNet

Selected DNNs

- VGG
- ResNet
- Inception

Paper Award.

Performance Loss

- CANN: Curable Approximations for Deep Learning Systems

MAC faults, Error bit from previous stage, Error bit from previous stage, Error bit from previous stage

The method offers 1.5x PDP reduction using the proposed MAC designs with no accuracy loss.