Towards a Bayesian Approach for Assessing the Fault Tolerance of Deep Neural Networks

Subho S. Banerjee, James Cyriac, Saurabh Jha, Zbigniew T. Kalbarczyk and Ravishankar K. Iyer
Computer Science, Electrical and Computer Engineering
Fault Injection Tools in Neural Networks

• Fault injection is one of the primary methods for assessing reliability validation/assessment

• Fault injection in NNs is difficult
  • Large space of fault locations and program states that must be injected/investigated
  • Need for significant system support to build system-specific injectors
  • Inability to provide statistical guarantees

• Question: Can we address these challenges by taking cognizance of latest developments in the machine learning space dealing with deep learning?
Bayesian Deep Learning & Fault Modelling

Deep Learning is compositions of functions on matrices.

\[ X \cdot w_1 + b_1 = Z' \]
\[ tanh(Z') = Z \]
\[ Z \cdot w_2 + b_2 = Y' \]
\[ relu(Y') = Y \]

Bayesian deep learning is composition of functions on probability distribution of matrices.

\[ X \cdot w_1 + b_1 = Z' \]
\[ Z' \cdot w_2 + b_2 = Y' \]
\[ relu(Y') = Y \]

Goal: Encode the fault model as a set of probability distributions over the parameters.
BDLFI: Bayesian Deep Learning Fault Injector

1. Neural Network
   - $b_i \sim \text{Bernoulli}(p) \quad \forall \ i \in [1, 32]$ 
   - $p$ based on AVF
   - $e = \{b_i\}$
   - $W' = e \oplus W$
   - Weights learned from golden run
   - $y' = \max(0, W'^T x + b')$

2. Bayesian Network based Failure Model
   - $E = \|W' - W\|$
   - Input from previous layer

3. Bayesian Inference
   - Original Classification Boundary
   - log(Error) Probability Due to Faults

   - PDF
     - $\text{flip}_p = 1e-05$
     - $\text{flip}_p = 0.061$
     - $\text{flip}_p = 0.1$
Looking Forward

• What advantage does this method give us?

• **Case 1**: Algorithmic Acceleration: *Fault injection == Monte Carlo*
  • Gradient-based Monte Carlo methods (NUTS sampler)
  • Importance sampling

• **Case 2**: Automate Reliable DL: *AutoML/Neural Architecture Search*
  • Design space: Duplication, TMR...
  • Approximations: Voltage scaling, DRAM refresh rate