Colony of NPUs
Babak Zamirai, Daya S Khudia, Mehrzad Samadi, Scott Mahlke

Approximate computing methodologies, Ultra-low cost mechanisms for fault and variability tolerant architectures

Approximate Computing
- Trade accuracy for
  - Performance
  - Energy consumption
- Neural Processing Unit (NPU):
  [Esmaeilzadeh et. al., Micro, 2012]

Better Accuracy is Expensive
- More accurate NPU need more
  - Hidden layers
  - Neurons per each layer
  - Memory accesses
  - Computation

Boosting Algorithms
- A set of weak learners create a strong learner
  - Replace a monolithic NPU with a set of small specialized NPUs
    - Improve accuracy
    - Reduce cost
    - Increase performance

Main Parts of cNPU
- Selector
  - Categorize error based on input
  - Decision tree
- Combiner
  - Trivial

Error Distribution
- Split
  - Main part: Most of the input data cause small errors
  - Tail part: Some inputs result in large errors
- Two NPUs
  - NPU_m: Most of the data will be sent to this simple NPU
  - NPU_t: A more complicated NPU to reduce that big errors

CNPU Configuration
- Minimize cost
- Minimize error

Minimize Cost
- Up to 95% cost reduction (same accuracy)
- 60% cost reduction on average without increasing error

Minimize Error
- Up to 75% error reduction (same cost)
- 31% error reduction on average without increasing cost

Future Work
- Reduce misclassification rate of selector
  - Design more accurate light-weight selectors
  - Utilize online feedback to tune accuracy of selector
- Divide input space based on features other than error
- Make small NPUs less power hungry
  - Neuron discarding
  - Synapse discarding
  - Precision reduction