**Input-Specialized Heterogeneous Neural Networks**

Babak Zamirai, Jiecao Yu, Salar Latifi, Scott Mahlke

---

**Introduction**
- More accurate neural networks need more
  - Hidden layers
  - Neurons per each layer
  - Memory accesses
  - Computation
- Neural network ensembles
  - A set of weak learners create a strong learner
  - Increases accuracy
  - All learners are active for each invocation
  - Energy-hungry
- Heterogeneity
  - Traditionally, each processor is specialized for a particular task
  - Data heterogeneity
    - Specialize each network for a part of training set
    - One active network per invocation, so consumes less energy
    - Each network focuses on a subset of data, so more accurate

---

**Data Heterogeneity**
- Main part: Small errors, dense
- Tail part: Large errors, sparse
- Partition training set systematically
- Use a specialized MLP (SMLP) for each part

---

**Multilayer Perceptrons**
- Feedforward fully-connected neural network
- Challenges
  - Model underfitting
  - High implementation cost
  - Accuracy limitation

---

**Multis-SMLP System**
- Replace a monolithic MLP with a set of SMLPs
  - Higher accuracy
  - Lower cost
  - Better performance
- Selector
  - Predict error during runtime
  - Always active
  - Decision tree

---

**Result**
- Beats boosting in terms of accuracy and energy
- Minimize error
  - Minimize energy

---

**Minimize Error**
- Reduce error by 66%
- Consuming 46% of the baseline energy

---

**Neural Network Ensembles**
- Replicating the baseline network
- Random sampling the training set
- Voting or weighted sum at the end
- Increase the overall energy multiplicatively

---

**Heterogeneous Neural Processing Unit**
- Hardware implementation of a 2-way multi-SMLP system

---

**More than Two SMLPs**
- Needs more complicated decision tree
- Increases the number of desirable configurations
- Does not improve accuracy and efficiency considerably