**Introduction**
- Complex neural networks need more
  - Hidden layers
  - Neurons per layer
  - Memory accesses
  - Computation
- Neural network ensembles
  - A set of weak learners creates a strong learner
  - Increases accuracy
  - Energy-hungry; all learners are active for each invocation
- 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
  - Improve energy efficiency and performance

**Background**
- Feedforward fully-connected neural network
- Challenges
  - Monolithic
  - Ensemble
  - Implementation cost
  - Number of NNs

**Motivation**
- Different subsets of critical neurons
- Monolithic NN is overdesigned
- Partition datasets and design specialized NNs

**Results**
- For 100% accuracy of baseline, reduce energy to
  - 32% on ARM Cortex-M4 Microcontroller
  - 38% on Kryo 280 octa-core Mobile CPU
  - 35% on Adreno 540 Mobile GPU

**Multi-dsNN System**
- Replace a monolithic NN with a set of dsNNs
  - Lower energy consumption
  - Higher performance
  - Selectors
    - Predict error during runtime
    - Always active
    - Decision tree

**Comparison with Boosting**
- Energy conscious boosting: limit energy budget
- Multi-dsNN provides the most energy-efficient design

**Partitioning Quality**
- Huge input layer
  - The decision tree might not work well
  - Complex dsNNs to maintain accuracy

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