Colony of NPUs: Scaling the Efficiency of Neural Accelerators

Babak Zamirai, Daya S Khudia, Mehrzad Samadi, and Scott Mahlke

University of Michigan, Ann Arbor

June 2015
Neural Processing Unit (NPU)

✦ Trade accuracy for

- Performance
- Energy consumption

[Esmaeilzadeh et. al., Micro, 2012]
Better Accuracy is Expensive

![Plot showing cost vs error with a red box highlighting the expensive region.](image-url)
Boosting Algorithms

✦ A set of weak learners create a strong learner

NPU1  NPU2

NPU3  NPU4

Big NPU
Colony of NPUs (cNPU)
Error Distribution

Split

- Main part
- Tail part
Main Parts of cNPU

✧ **Selector**
  - Categorize error based on input
  - Decision tree
  - Tiny neural network

✧ **Combiner**
  - Trivial
Training cNPU

Exact output

Input

Big NPU

Comparator

Error

Selector

NPU⁺

NPU⁻

Selector

Runtime

Input

Output
Opportunity of Improvement

✧ Different configurations for NPU\textsubscript{m} and NPU\textsubscript{t}
✧ Different splitting thresholds

K-Means (2,2)  Black-Scholes (8,8)
cNPU Configuration

- Reduce cost
- Improve accuracy

**K-Means (2,2)**

- $NPU_t = (2,1)$
- $NPU_m = (1,2)$
- Threshold = 0.17

**Black-Scholes (8,8)**

- $NPU_t = (3,2)$
- $NPU_m = (2,1)$
- Threshold = 0.21

- $NPU_t = (7,8)$
- $NPU_m = (7,8)$
- Threshold = 0.16
Evaluation

Minimizing Cost
Conclusion

- Better accuracy need bigger NPUs
- cNPU is combination of small NPUs
- Up to 87% cost reduction (same accuracy)
- Up to 1.95x accuracy improvement (same cost)