Pairing Up CNNs for High Throughput Deep Learning
Babak Zamirai, Salar Latifi, Scott Mahlke

DNN Complexity Trend
- Computational complexity grows fast
  - Accuracy improvement
  - Input-invariant accelerations

Input Variation
- There is no single best CNN for all inputs
- Combine multiple CNNs
  - Lower computational complexity
  - Higher accuracy

Odd Correct

Common Correct

Complex Correct

ResNet 152
54%
79%

AlexNet
57%

DNN Complexity Trend

Runtime
- Run everything on the little CNN
- Detect and recover unreliable outputs

Softmax Layer
- An estimation of confidence
- Sum of the elements = 1.0

Normalized output

Confidence Probe
- Recovery rate = 26%
- Odd corrects maintain the accuracy

Synergistic Pairs
- Higher accuracy means more room for savings
- Odd corrects and peak accuracy are correlated
- More odd corrects, better synergy

Latency
- Exhaustive search results in only 5% additional gains

Datacenter Throughput
- Same response time as baseline