**Bit-Tactical**: Exploiting Ineffectual Computations in Convolutional Neural Networks: Which, Why, and How

Alberto Delmas, Patrick Judd, Dylan Malone Stuart, Zissis Poulos, Mostafa Mahmoud, Sayeh Sharify, Milos Nikolic, Andreas Moshovos

**Abstract**

**Bit-Tactical**: Deep neural network inference accelerator

- Targets primarily CNNs, can do any layer type
- Exploits sparsity and effectual bit content

Over **10x** faster than data-parallel accelerators

2x more energy efficient

32% area cost

**Motivation**

Previous accelerators exploit data sparsity in one dimension (zero-values, precision variability, zero-bits; on either Weights or Activations). Combining zero-value Weight sparsity along with zero-bit Activation sparsity exploitation provides the best performance potential for sparse networks, while still benefiting dense ones.

![Graph showing performance improvement](image)

**Most computation is ineffectual: over 31x potential**

**Baseline**

Straight-forward implementation

Synchronized parallel units that do not skip any work:

![Diagram of baseline implementation](image)

**Backend: precision adaptability**

Over 90% of Activation bits are **ZERO**

![Diagram of backend adaptability](image)

**Off-Chip bandwidth characterization**

![Diagram of off-chip bandwidth](image)

**Results**

**Throughput**

![Graph showing throughput comparison](image)

**Energy efficiency**

![Graph showing energy efficiency](image)

**Comparison with other accelerators**

![Graph showing comparison with other accelerators](image)

**Organization**

![Diagram of organization](image)