MOODY: Data-Aware Reconfiguration for Energy Efficient Deep Learning

YArch HPCA 2019
Pedram Zamirai, Armand Behroozi
Advisor: Scott Mahlke
Diversity in DNNs

Diversity in DNNs

- Fully Connected (FC)
- Convolutional (Conv)
Static/Dynamic

![Diagram of static and dynamic processes]

- Profile
- Run
MOODY: Dynamic/Dynamic
MOODY: Dynamic/Dynamic
Precision Reduction For Training

Full precision activations

Conv1  Conv2  Conv3
Precision Reduction For Training

Needed bits

Redundant bits

Conv1

Conv2

Conv3
Precision Reduction For Training

Operate on fewer bits!

Conv1

Conv2

Conv3
Activation Density During Training

Density is dynamic!

GoogLeNet

Activation Density During Training

Training Time
Summary

- MOODY: Reconfigurable data-aware accelerator
  - Lower precision computation
  - Activation density changes during training
- Improves both performance and energy
  - Higher effective memory bandwidth
  - Higher compute bandwidth
  - Increased overall throughput
MOODY: Data-Aware Reconfiguration for Energy Efficient Deep Learning

YArch HPCA 2019
Pedram Zamirai, Armand Behroozi
Advisor: Scott Mahlke