Scaling Deep Neural Network Accelerator Performance

CMC AI Workshop

Pierre Paulin, Director of R&D
18 February 2020
Outline

• Deep Neural Network Trends
• EV7x Processor and DNN Engine Overview
  – Specialized DNN accelerator
  – Local optimization of data movement
  – Local data compression of coefficient and feature-maps
• Advanced Bandwidth Optimization Techniques
  – DMA broadcast of coefficients and feature-maps
  – Multi-level layer fusion
  – Multi-level tiling across memory hierarchy
Deep Neural Network Trends
Trends in Convolutional Neural Networks Topologies

Trend 1: Reduced Computational Requirements
Trend 2: Reduced Model Size
Trend 3: Reduced Data Reuse and Parallelism
Trend 4: Feature-map Bandwidth Becomes Dominant

Examples: MobileNet, DenseNet
Trend 1: Reduced Computational Requirements

Nearly 100X reduction
Trend 2: Reduced Model Sizes

- 2018

- % Accuracy Top 1
- Milion Weights

Graph showing the relationship between model sizes and accuracy top 1 for various models:
- AlexNet
- VGG16
- Inception
- ResNet
- Mobilenet V1
- Mobilenet V2
Trend 3: Reduced Data Reuse and Parallelism

Example: Depthwise Separable Kernels used in MobileNet V2

- **Traditional 1x1 Convolution**
  - Conv 1x1: 64 → 256
  - DW Conv 3x3: 256
  - Conv 1x1: 64 → 64
  - High Computation
  - High Data Reuse
  - High Parallelism

- **Depth-wise Separable 3x3 Convolution**
  - Conv 1x1: 64
  - +
  - 64
  - Low Computation
  - Low Data Reuse
  - Low Parallelism
Trend 4: Feature-map Bandwidth Becomes Dominant

Example: DenseNet and Multilayer DenseNet

More Connections between Layers

→ More Bandwidth for Feature-maps
Trends in Convolutional Neural Networks Topologies

Trend 1: Reduced Computational Requirements
Trend 2: Reduced Model Size
Trend 3: Reduced Data Reuse and Parallelism
Trend 4: Feature-map Bandwidth Becomes Dominant

Examples: MobileNet, DenseNet
Scaling Performance with Bandwidth Constraints

- Bandwidth reduction has direct impact on performance and power
- Over 50% of SoC power is DRAM access
Embedded Vision Processor Outline

• Deep Neural Network Trends
  – Accuracy and Functionality

• **EV7x Processor Family Overview**

• DNN Engine
  – Specialized DNN accelerator
  – Local optimization of data movement
  – Local data compression of coefficient and feature-maps

• Advanced Bandwidth Optimization Techniques
  – DMA broadcast of coefficients and feature-maps
  – Multi-level layer fusion
  – Multi-level tiling across memory hierarchy
EV7x Processor and DNN Engine Overview
EV7x Vision Processor IP with 35 TOPS Performance

14,080 MAC Engine Made Possible with Better Utilization, Bandwidth & Power

- **Addresses market requirements** for full range of vision applications: always-on IoT, augmented reality, autonomous driving…
- **Faster neural network accelerator** executes all graphs include the latest, most complex graphs
- **Enhanced vision engine** for low-power, high-performance Vision, Simultaneous Localization and Mapping (SLAM) and DSP algorithms
- Architectural changes and power gating techniques **reduce power consumption**
- **High-bandwidth encryption** protects coefficients and biometric data
- **Automatic graph partitioning** using MetaWare EV for improved performance, bandwidth, latency
EV6x/7x Scalable DNN Engine for Deep Learning-based Vision
- High performance, low power and low area
- Fully programmable
DNN Accelerator Supports Up to 35 TOPS For All DNN Applications

0.1 to 35 TOPS to Address All Vision Applications

• Deep Neural Network Engine supports
  – Convolutional Neural Networks (CNN)
  – Batched Recurrent Neural Networks (RNN)

• EV7x max performance
  – Up to 14,080 multiply-accumulators per engine

• Improved utilization provide increases MAC efficiency
  – Higher MAC utilization for 1x1 and 3x3 convolutions
  – Increased support for non-linear functions (PReLU, ReLU6, Maxout, Sigmoid, Tanh, …)

• Architectural enhancements improve bandwidth, accuracy and power
Graph Mapping: Support of Multiple CNN Frameworks

- Support new graph frameworks via ONNX-based interoperability
  - ONNX export utilities being made available for numerous frameworks
- Neutral Intermediate representation
  - Integrates the union of Caffe, Tensorflow, ONNX features
The Bandwidth Challenge
Scaling Performance with Bandwidth Constraint

- Bandwidth reduction has direct impact on performance and power
- Over 50% of SoC power is DRAM access
EV DNN Bandwidth Improvement Features

• Coefficient Pruning
  – Coefficients with a zero value are skipped/counted

• Feature Map Compression
  – Lossless runtime compression and decompression of feature maps to external memory

• Multi-level Layer Fusion
  – Merging multiple folded layers into single primitives reduces feature map bandwidth

• Optimized Handling of Coeff. and Feature-maps
  – Sharing of common data across slices to minimize bandwidth of coefficients and feature-maps loading
Feature-map partitioning

Split each layer over multiple slices

- Higher throughput – up to 4X
- Lower latency – up to 4X – due to parallel processing of a layer
- Significant bandwidth reduction
• Opposing CNN Graph Trends
  – Reduced compute requirements and model size
  – Reduced data reuse and parallelism
  – Feature-map bandwidth becomes dominant
• Synopsys DNN Engine
  – Specialized DNN accelerator
  – Local optimization of data movement
  – Local data compression of coeff. and feature-maps
• Advanced Bandwidth Optimization Techniques
  – Optimized handling of coefficients and feature-maps
  – Multi-level layer fusion and tiling
• Improved scalability, lower power
  – 10 TOPs/W (7 nm)
Thank You