“The Cost of Everything and the Value of Nothing”

David Moloney, CTO

Hot Chips 28, Cupertino, California
August 23, 2016
DEEP LEARNING: THE GREAT DISRUPTOR

“Tesla Zooms Past BMW, Audi Limos In Europe, Closes In On Mercedes"
AGENDA

1. Deep Learning: The Great Disruptor
2. Training vs Inference: Inference Matters in Embedded
3. The Deeper the Better?
4. Benefits of Embedded Processing at Network Edge
5. Maximising Performance of Networks on Myriad 2 MA2x50
6. Optimising CNNs on Vector Processors for Mobile Platforms
7. Conclusions
DEEP CONVOLUTIONAL NEURAL NETWORKS (CNNs)

Inference Matters in Embedded

Layer 1
Layer 2
Layer 3
Layer 4
Layer N
Output
Input
Weights

Training
Error
+

Backpropagation

Input
Weights

Expected Output (label)

Layer 1
Layer 2
Layer 3
Layer 4
Layer N
Output

Convolution
Max-Pooling (blockwise maximum)
Activation (pointwise ReLU)

Recurrent for N layers in Network

Expected Output (label)

Weights

Inputs Previous layer or DDR/RAM

Inputs to next layer

+
AGENDA

- Deep Learning: The Great Disruptor
- Training vs Inference: Inference Matters in Embedded
- The Deeper the Better?
- Benefits of Embedded Processing at Network Edge
- Maximising Performance of Networks on Myriad 2 MA2x50
- Optimising CNNs on Vector Processors for Mobile Platforms
- Conclusions
THE DEEPER THE BETTER…
Yet With Increased Complexity

ImageNet Accuracy % over Time

Year

<table>
<thead>
<tr>
<th>Year</th>
<th>Layers</th>
<th>Accuracy %</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>8</td>
<td>83.60%</td>
</tr>
<tr>
<td>2013</td>
<td>16</td>
<td>88.80%</td>
</tr>
<tr>
<td>2014</td>
<td>58</td>
<td>93.33%</td>
</tr>
<tr>
<td>2015</td>
<td>158</td>
<td>95.06%</td>
</tr>
</tbody>
</table>

- AlexNet
- VGG16
- GoogLeNet
- ResNet
COMPLEXITY OF A STANDARD NETWORK

GoogLeNet

- Computational cost is increased by less than 2x compared to AlexNet (<1.5 Bn operations/evaluation)
- 5M parameters
AGENDA

- DEEP LEARNING: THE GREAT DISRUPTOR
- TRAINING VS INFERENCEx: INFERENCEx MATTERS IN EMBEDDED
- THE DEEPER THE BETTER?
- BENEFITS OF EMBEDDED PROCESSING AT NETWORK EDGE
- MAXIMISING PERFORMANCE OF NETWORKS ON MYRIAD 2 MA2x50
- OPTIMISING CNNs ON VECTOR PROCESSORS FOR MOBILE PLATFORMS
- CONCLUSIONS
EMBEDDED PROCESSING AT THE NETWORK EDGE

Huge Benefits

- 1,000,000 x more energy-efficient
- 10,000 x less bandwidth consumed
- 1,000 x lower latency

Other Improvements in:
- Privacy
- Fault-tolerance/ Continuity of service
AGENDA

DEEP LEARNING: THE GREAT DISRUPTOR

TRAINING VS INFEERENCE: INFEERENCE MATTERS IN EMBEDDED

THE DEEPER THE BETTER?

BENEFITS OF EMBEDDED PROCESSING AT NETWORK EDGE

MAXIMISING PERFORMANCE OF NETWORKS ON MYRIAD 2 MA2x50

OPTIMISING CNNs ON VECTOR PROCESSORS FOR MOBILE PLATFORMS

CONCLUSIONS
MYRIAD 2 MA2x50
Vision Processing Unit (VPU) Architecture

Software Controlled I/O Multiplexing

AMC Crossbar

CMX Memory Fabric Multi-Ported RAM Subsystem

Inter-SHAVE Interconnect (ISI)

Arbiter & 16:1 mux
L2 cache 256KB
DDR Controller

Stacked KGD 1-8Gbit LP-DDR2/3

>20 independent power islands

PLL & CPM
SHAVE PROCESSING
Maximising GEMM Performance

Computes 2 rows of $A^*B = 2$ rows of $C$
16 lines of SHAVE VLIW Assembler

8-bit MACs in VAU & 32b Partial Summation $A0n*B0n n\{0\cdots7\}$ in SAU
32-bit Accumulation in IAU

Zero Overhead Looping

Load/Store Port 0
Load/Store Port 1
## GoogLeNet RELATIVE GFLOPS/W

### Performance Results

**GoogLeNet Single Inference (Batch = 1) no Heatsink or Fan**

<table>
<thead>
<tr>
<th>nm Process</th>
<th>fps</th>
<th>Power (W)</th>
<th>GFLOPS</th>
<th>GFLOPS/W</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPU-A</td>
<td>28</td>
<td>15</td>
<td>7.5</td>
<td>51.37</td>
</tr>
<tr>
<td>GPU-B</td>
<td>20</td>
<td>22</td>
<td>7.5</td>
<td>75.34</td>
</tr>
<tr>
<td>Myriad 2</td>
<td>28</td>
<td>25</td>
<td>1.2</td>
<td>85.61</td>
</tr>
</tbody>
</table>

- **nm Process**
- **fps**
- **Power (W)**
- **GFLOPS**
- **GFLOPS/W**
AGENDA

- DEEP LEARNING: THE GREAT DISRUPTOR
- TRAINING VS INFEERENCE: INFEERENCE MATTERS IN EMBEDDED
- THE DEEPER THE BETTER?
- BENEFITS OF EMBEDDED PROCESSING AT NETWORK EDGE
- MAXIMISING PERFORMANCE OF NETWORKS ON MYRIAD 2 MA2x50
- OPTIMISING CNNs ON VECTOR PROCESSORS FOR MOBILE PLATFORMS
- CONCLUSIONS
HOW MOVIDIUS IS DEPLOYING STANDARD CNNs
At the Network Edge

CNN Model
Description

Weights

For Integration

Efficient Execution of
Kernel Function Libraries

PARSES CNN Model

Caffe
TensorFlow

DRONES AR/VR
ROBOTICS SECURITY

www.movidius.com
© Copyright Movidius 2016
AGENDA

1. DEEP LEARNING: THE GREAT DISRUPTOR
2. TRAINING VS INFEERENCE: INFEERENCE MATTERS IN EMBEDDED
3. THE DEEPER THE BETTER?
4. BENEFITS OF EMBEDDED PROCESSING AT NETWORK EDGE
5. MAXIMISING PERFORMANCE OF NETWORKS ON MYRIAD 2 MA2x50
6. OPTIMISING CNNs ON VECTOR PROCESSORS FOR MOBILE PLATFORMS
7. CONCLUSIONS
IS IT WORTH IT?
Incremental Accuracy and the Power Efficiency Cost

Notes: ImageNet, Batch = 10/64, using active cooling
Deep Learning for Embedded is all about **Inference**

Standard Networks are designed to achieve **high-accuracy**

Embedded implementation on architectures such as Movidius VPU can achieve **significant performance results** at the network edge

Next challenge is to further optimise networks to **maximise performance per Watt**

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement n°611183 (EXCESS Project, www.excess-project.eu).
THANK YOU
FOR YOUR ATTENTION