Power Efficient Architectures to Accelerate Deep Convolutional Neural Networks for edge computing and IoT

Giuseppe Desoli

ST Central Labs
STMicroelectronics
Artificial Intelligence is Everywhere

• **Analysis**, i.e.:
  • Where am I?
  • Scene classification (audio, video, environmental sensors)
  • Which objects are in the scene, where are they?
  • Video object detection/classification
  • What am I doing?
  • Activity recognition (audio, video, inertial sensors)
  • What’s happening?
  • Event recognition (audio, video, inertial sensors, environmental sensors).

• **User Interaction**:  
  • Command detection (audio)
  • Speech Recognition (audio)
  • Gesture Recognition (inertial sensors, video)
  • User identification and mood detection (audio, video)

• **Continuous Learning**, i.e.:
  • How can I detect unpredictable, unclassified events in dynamic environments?
  • Recurrent networks (audio, video, inertial sens, environm sens)

• And many more…..
Distributed Intelligence: why?

Scalability
Responsiveness
Smart units

Big data
Service enablement
Global optimization

Collected data analytic
Service enablement

Local analytic
Real time

Centralized
Distributed

100 sensors
100 Mb-1 Tb / sec
10 TOPS-10000 TOPS

10000 sensors
~10 Mb / sec
~10 GOPS

100 sensors
1 Mb-10 Gb / sec
100 MOPS-100 TOPS

100 sensors
~100 Kb / sec
~100 MOPS

1 sensor
10 KOPS-100 MOPS

1 Intelligent sensor
100 MOPS-1 TOPS

1 sensor
10 Kb-100 Mb / sec

“Artificial neural network”
placing intelligence
where required

“Artificial neural network”
placing intelligence
where required

10000 sensors
100 Mb-1 Tb / sec
10000 TOPS

10000 sensors
~10 Mb / sec
~10 GOPS

100 sensors
1 Mb-10 Gb / sec
100 MOPS-100 TOPS

100 sensors
~100 Kb / sec
~100 MOPS

1 sensor
10 KOPS-100 MOPS

1 Intelligent sensor
100 MOPS-1 TOPS

1 sensor
10 Kb-100 Mb / sec
Why Artificial Neural Networks?

- The power and usefulness of ANN have been demonstrated in several applications.
- A specific kind of neural networks, the Deep Convolutional Neural Networks (DCNN), have proven very effective.
  - Achieving human-like performance in selected cases.
- In the last decade breakthroughs made neural networks practical.
  - Better training algorithms, Moore’s law and Big data availability.
- For IoT the current challenge is to achieve low power and adequate cost with sufficient performance for edge computing.
A Typical CDNN Structure

- The Artificial Neuron is a ‘processing unit’ with a close connection to neurobiology
- DCNN’s are composed of **multiple layers** of ‘neurons’
- Each layer performs **feature extraction** with ‘learned’ filters, **reduction** of input resolution and **non-linear** operations
- Multiple layers ‘**compress**’ each image into denser information
- Depth indicates the **number of layers** of the specific CDNN network.
- **Up to millions of parameters** for each layer of the network can be involved
- Parameters can be defined thanks to supervised or unsupervised training algorithms processing **large training sets**
Deep Learning on images

- Image Classification
- Object Localization
- Object Detection
- Image Segmentation
- Action Recognition
- Image Generation

Images source: ImageNet
Beyond recognition: semantic captioning

- a giraffe has its head up to a small tree.
- a giraffe in a pen standing under a tree.
- giraffe standing next to a wooden tree-like structure.
- a tall giraffe standing next to a tree
- a giraffe in an enclosure standing next to a tree.

Courtesy of COCO: Common Object in Context, Microsoft
Deep Learning for Speech

- Speech Recognition
- Natural Language Processing
- Speech Translation
- Audio Generation

The new WaveNet model is 1,000 times faster and can produce 20 seconds of higher quality audio in 1 second.

https://www.flickr.com/photos/tevk/5429390495/
Deep Learning for autonomous driving

- Simultaneous objects (cars, pedestrian, signals) detection and identification
- Semantic segmentation
- Multiple sensory input (visual, radar, lidar, proximity, etc.)
- End to end processing and actuation
Convolutional NN Complexity Evolution

- **Operations (GOPS)**
- **Parameters (Millions)**

<table>
<thead>
<tr>
<th>Model</th>
<th>Year(s)</th>
<th>Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANNs</td>
<td>1997-2007</td>
<td>3 layers</td>
</tr>
<tr>
<td>AlexNET</td>
<td>2012</td>
<td>8 layers</td>
</tr>
<tr>
<td>GoogleLeNet</td>
<td>2014</td>
<td>22 layers</td>
</tr>
<tr>
<td>VGG19</td>
<td>2014</td>
<td>19 layers</td>
</tr>
<tr>
<td>ResNet</td>
<td>2015</td>
<td>152 layers</td>
</tr>
</tbody>
</table>
HW for Deep Learning: few examples

Intel Xeon Phi 7285
Freq 1.4GHz, 68 cores,
TDP 250W, 14nm,
Perf > 3.4TOPS (TBA)
Price $2036 (sept 2017)
instructions for deep learning
(AVX512-4VNNIW, AVX512-4FMAPS)

NVIDIA Xavier
8-core CPU,
512-core Volta GPU
30 TOPS
TDP 30W, 16nm

Mobileye (now Intel)
EyeQ5
7nm
12 TOPS peak
2.4 DL TOPS

Movidius Myriad X
(now Intel)
16 vector 128bit VLIW,
Neural Compute Engine
4 TOPS
TDP 2W, 16nm
Exploiting parallelism (we need special HW)

- Two broad classes of architectures can be identified
- Both have pros and cons
- Specialized HW is needed to achieve power consumption compatible with IoT applications and cost
- Memory access the key aspect

Temporal Architectures (SIMD/SIMT)

Spatial (Data-flow) Architectures (SIMD)

Energy/power x word access

- Local SRAM: 1x
- On-chip SRAM: 10x
- LPDDR: 100x

Courtesy of MIT Eyeriss project
An Ultra low power example: Orlando SoC
Reconfigurable Accelerator Framework

Virtual stream links

- Ferry data to/from accelerators, interfaces and DMA engines
- Flow control mechanism is provided
- Streams can be multicast to multiple destinations

- More flexible than hardware data paths
- More power efficient than a bus
**Prototype Chip FD-SOI 28nm**

<table>
<thead>
<tr>
<th>Technology</th>
<th>FD-SOI 28nm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Package</td>
<td>FBGA 15x15x1.83</td>
</tr>
<tr>
<td>Frequency</td>
<td>200MHz–1.175GHz</td>
</tr>
<tr>
<td>Supply voltages</td>
<td>0.575–1.1 V digital 1.8V I/O</td>
</tr>
<tr>
<td>Power (**)</td>
<td>41 mW @ 42 FPS</td>
</tr>
<tr>
<td>@200Mhz, 0.575V, 8 CAs</td>
<td></td>
</tr>
<tr>
<td>On-chip RAM</td>
<td>4x1MB 8x192KB 128KB</td>
</tr>
<tr>
<td>Host</td>
<td>ARM® Cortex®-M4</td>
</tr>
<tr>
<td>No of DSPs</td>
<td>16</td>
</tr>
<tr>
<td>Peak DSP perf</td>
<td>75 GOPS (2x16bMAC(*))</td>
</tr>
<tr>
<td>No of CAs</td>
<td>8</td>
</tr>
<tr>
<td>CAs perf (1.175GHz, 1.1V)</td>
<td>676 GOPS((^\ast)) peak</td>
</tr>
</tbody>
</table>

(*) 1 MAC defined as 2 OPS (ADD + MUL)
(**) HW Acc avg power for AlexNet
Ultra-Wide DVFS Range

- LVT design with heterogeneous Poly-Bias levels \(\Rightarrow\) perf vs leakage
- GALS and low insertion delay clock networks to minimize on chip variation margins;
- Mono Supply memories with fine grained power switches and sleep mode;
- DVFS energy efficiency improvements via body bias.
Application Example: AlexNet

Input image

37.5 mW @ 200MHz, 0.6V
10 FPS (38 ms DSPs, 62 ms CAs) 2 chained CAs
Dynamic: 10 mW CAs + 17 mW system
Static: 0.6 mW CAs + 9.9 mW system
Orlando CNN inference engine performance

Compared (unfairly) to NVIDIA Tegra X1 @
1200 MHz FP32
83 FPS
Cost: 500-1000$
TDP: > 200W
Orlando at work

- Left: Orlando running Pico Yolo CNN for object detection and classification
- Top: Orlando running a CNN trained to drive a simulated car
- Bottom: Orlando identifying faces and classifying expressions