Efficient Implementation of Neural Networks

Jeff VanWashenova – Director of Automotive Market Segment
Presentation Outline

1. CEVA
2. Automotive Trends Challenges in Deep Learning
3. Efficient implementation of Neural Networks
4. Alexnet Example
CEVA – Ultra Low Power IP

- Image and Vision, Connectivity, Communication
- +7 Billion Devices

CEVA

Silicon

Tier 1s
Potential Tier 1s such as:
Bosch, Continental,
Delphi, Valeo, etc.

Silicon IP

Chips

Development Tools, Support, SW

OEM
Potential OEMs such as:
GM, Ford, BMW, Toyota, Mazda, Daimler, …

Systems, SW
CEVA Vision Platform

- Platform includes vision processor, libraries, tools and applications (CEVA and a variety of SW partners and service experts)

- 4th-generation imaging and vision processor IP

- Vector-type processor; combines fixed- and floating-point math; up to 4096-bit processing per cycle

- XM4: 10+ design wins, Silicon available in Q2/2016

- Platform enables efficient embedded vision applications
Presentation Outline

1. CEVA
2. Automotive Trends Challenges
3. Efficient implementation of Neural Networks
4. Alexnet Example
Deep Learning in Automotive
High volume vision applications leading to autonomous driving

- Neural Networks
  1. Optimal for harsh automotive environment – Covered signs
  2. Re-trainable without code changes (implement once, use many times)

- Explosion of computational load; fixed power budget

- Challenges
  - **Cost** – Need for cost efficient embedded systems
  - **Power Efficiency** – Fixed power budget – 10 – 15W Centralized
  - **Time to Market** - Migration to production
Presentation Outline

1. CEVA
2. Automotive Trends Challenges
3. Efficient implementation of Neural Networks
4. Alexnet Example
Efficient Implementation of Neural Networks

**Cost**
- **Specialization**: IP that is optimized for vision and deep learning
- **Die Size**: Specialized core for vision and neural networks are smaller than GPU for best performance/mm²

**Efficiency**
- **Utilization**: Dedicated Vision Processor can achieve higher utilization of core
  - > 95% utilization improves Performance/Watt/
  - **Floating point to Fixed Point Operations**
    - Fixed point sufficient for CNN

**Time to Market**
- **Flexibility**: Quickly adapt to algorithm and network changes
- **Qualifications**: Select designs with need safety packages (ISO26262)
- **Software Libraries**: Quick development and deployment of common vision/CNN functions
- **Software Tools**: SW Tools to migrate from R&D to Production quickly

- **17x** Smaller Die Size*
- **9x** Lower Power**
- **30%** Faster Processing*

* vs. GPU-based systems **vs. typical implementation
CNN Usage Flow with Caffe & CDNN

Network Structure → Caffe → Floating-point Network + Weight → CEVA Network Generator → Fixed-Point Customized Network + Weights

Training Stage (Offline) → Network Weights → Detection Stage (real-time) → "DOG"

Image Database
Real-Time CDNN Application Flow

Inputs
Application Pre-process: Scaling, background reduction, ROI selection

CDNN Real-Time Libraries

CDNN Application API
- Convolution API
- Convolution Layer
- Normalization API
- Normalization Layer
- Activation Layer API
- Neuron Activation Layer
- Pooling API
- Pooling Layer
- Fully Connected API
- Fully Connected Layer
- Soft Max API
- Soft Max Layer

Full Network Implementation

Application Post-process: Complete recognition, ...

"DOG"
Presentation Outline

1. CEVA
2. Automotive Trends Challenges
3. Efficient implementation of Neural Networks
4. Alexnet Example
Example CNN – AlexNet
Example based on Caffe open source implementation for CNN

<table>
<thead>
<tr>
<th>Object</th>
<th>AlexNet PC Probability (floating point)</th>
<th>AlexNet on XM4 Probability (fixed point)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labrador retriever</td>
<td>90.44%</td>
<td>91.01%</td>
</tr>
<tr>
<td>Golden retriever</td>
<td>4.45%</td>
<td>3.98%</td>
</tr>
<tr>
<td>Beagle</td>
<td>0.21%</td>
<td>0.18%</td>
</tr>
<tr>
<td>Kuvasz</td>
<td>0.12%</td>
<td>0.10%</td>
</tr>
</tbody>
</table>

Classification Probabilities
CEVA-XM4 CDNN Development Platform

i.MX6

Host running Linux applications

XM4 FPGA

PCIe
**CEVA-XM4 CDNN Demo**

- Live Alexnet object recognition – *come visit our booth!*
- Enables milli-watt products vs. watts on GPU

---

**Diagram Description**

- **Input Images**: Webcam FHD
- **CEVA Host**: iMX6
- **Memory**: Shared Memory
- **HDMI**: Daisy
- **USB**: Webcam FHD
- **PCIe**: CEVA Host Link
- **JBOX**: PC Debugger
- **XM4 FPGA**: DDR, DMA, Data TCM, Code TCM, Code Cache, CEVA Link, CDNN Engine
- **Conversion**: FHD to 224x224
- **Live**: Alexnet object recognition – come visit our booth!
CEVA – Efficient System Approach

**XM4 – Vision Processor**
- Flexible: Fully programmable
- Efficient: Near full utilization of core processing
- Fixed point and floating point operations
- Optimized for image and vision applications
- Small Die size to enable cost effective designs

**Offline Network Generator**
- Auto convert fully trained networks
- Floating to fixed point conversion – Embedded Target
- Network optimized for embedded vision IP core – CEVA XM4

**Real Time NN Libraries**
- RT Libraries available for quick algorithms development and deployment
- Optimized for CEVA-XM4 vision DSP
- Common vision and neural network libraries

CEVA Vision Platform approach for efficient implementation
Thank You