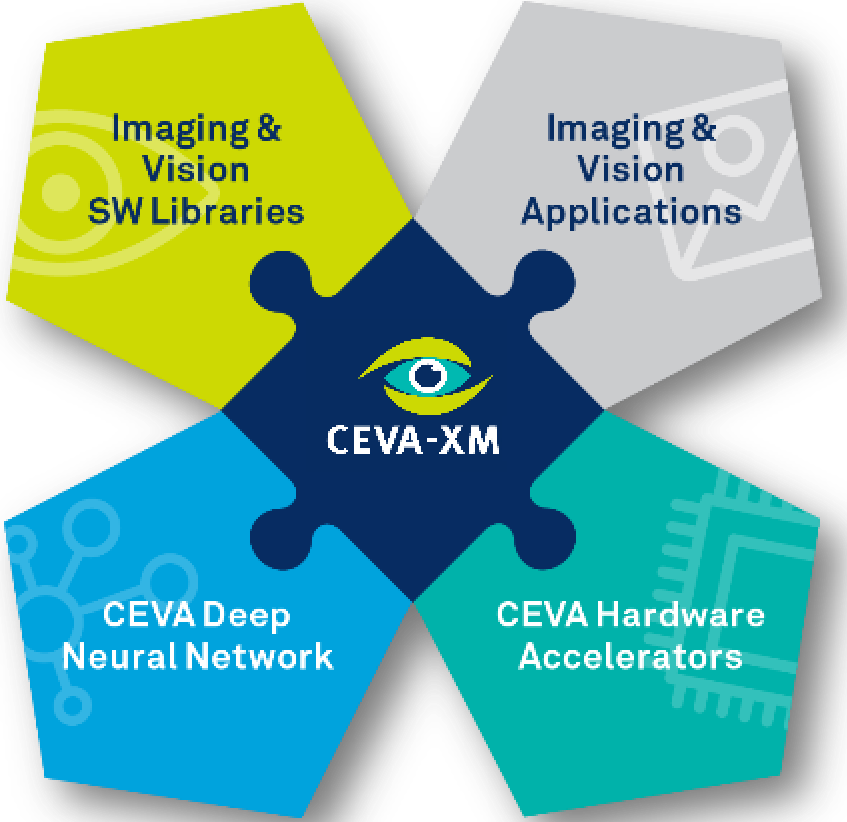
CEVA Deep Neural Network (CDNN) Introduction

May, 2017 – Under NDA

www.ceva-dsp.com

CEVA's Imaging & Vision Technology

- Comprehensive vision platform
- Centered on CEVA-XM Vision DSP
- Enables embedded neural networks for mass market intelligent vision applications
- Simplifies delivery of powerful deep learning solutions on low-power embedded devices



CEVA Imaging & Vision Market Adoption

- CEVA-**XM6**
  - 5th generation
  - 5+ design wins
- CEVA-**XM4**
  - 4th generation, in production
  - 30+ design wins
  - Available open vision DSP in the market
    - By Rockchip, Novatek and Brite Semi
- CEVA-**MM3101**
  - 3rd generation, in production
  - 20+ design wins
  - Available open vision DSP in the market
    - By Socionext, Inuitive and Novatek

CEVA Vision DSP Public Customers	
LG	Rockchip
NOVATEK	Panasonic
ON Semiconductor	altek
socionext	VIA
VATICS	INUITIVE
iCatch Technology, Inc.	Brite semiconductor

CEVA processors are de-facto standard for Imaging & Vision

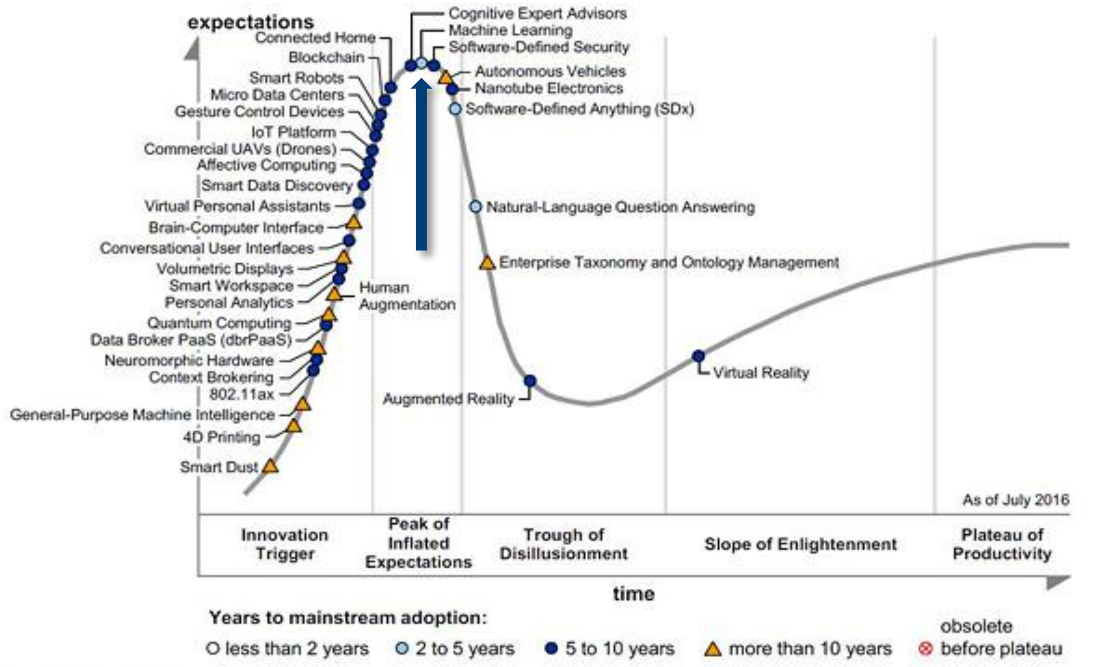
Outline

- Neural Network Introduction and Embedded Challenges
- CEVA Deep Neural Network (CDNN) Toolkit
- CDNN2 SW Framework
- CNN HWA
- CDNN Performance
- CDNN Roadmap

Hype Cycle for Emerging Technologies

CEVA®

2016: Machine Learning at the hype peak



Source: Gartner's Aug 2016 Hype Cycle for Emerging Technologies

Neural Network Embedded Challenges

Implementing a deep neural network in an embedded systems is an extremely challenging task!

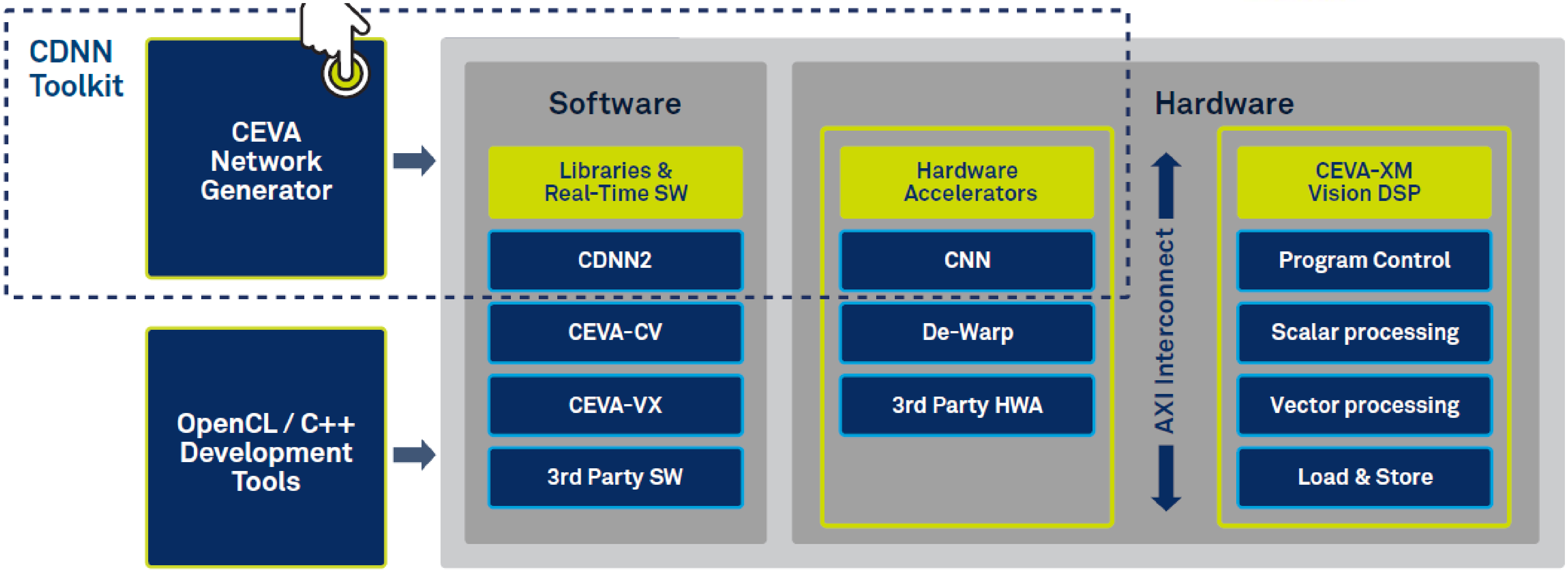
Very high bandwidth consuming and computing bottleneck

Porting and optimization capabilities

Stringent power budget and system cost

Long “Time-To-Market”

CEVA's Imaging & Vision Technology

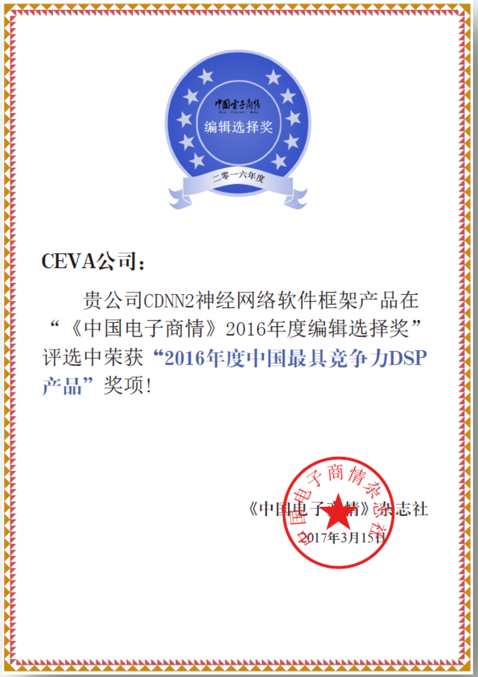


Comprehensive and Scalable Vision and Deep Learning Solution

CDNN2 – CEM 2016 Editor's Choice Awards

- **About China Electronic Market (CEM)**
  - Monthly magazine founded in 1995
  - Focus on electronics and semiconductors in China
  - Provides coverage of new products, technical and market trends, and market data
  - Supported by China’s Ministry of Industry and Information Technology (MIIT) and has a circulation of around 28,000

CDNN2



CEVA公司：

贵公司CDNN2神经网络软件框架产品在“《中国电子商情》2016年度编辑选择奖”评选中荣获“2016年度中国最具竞争力DSP产品”奖项！

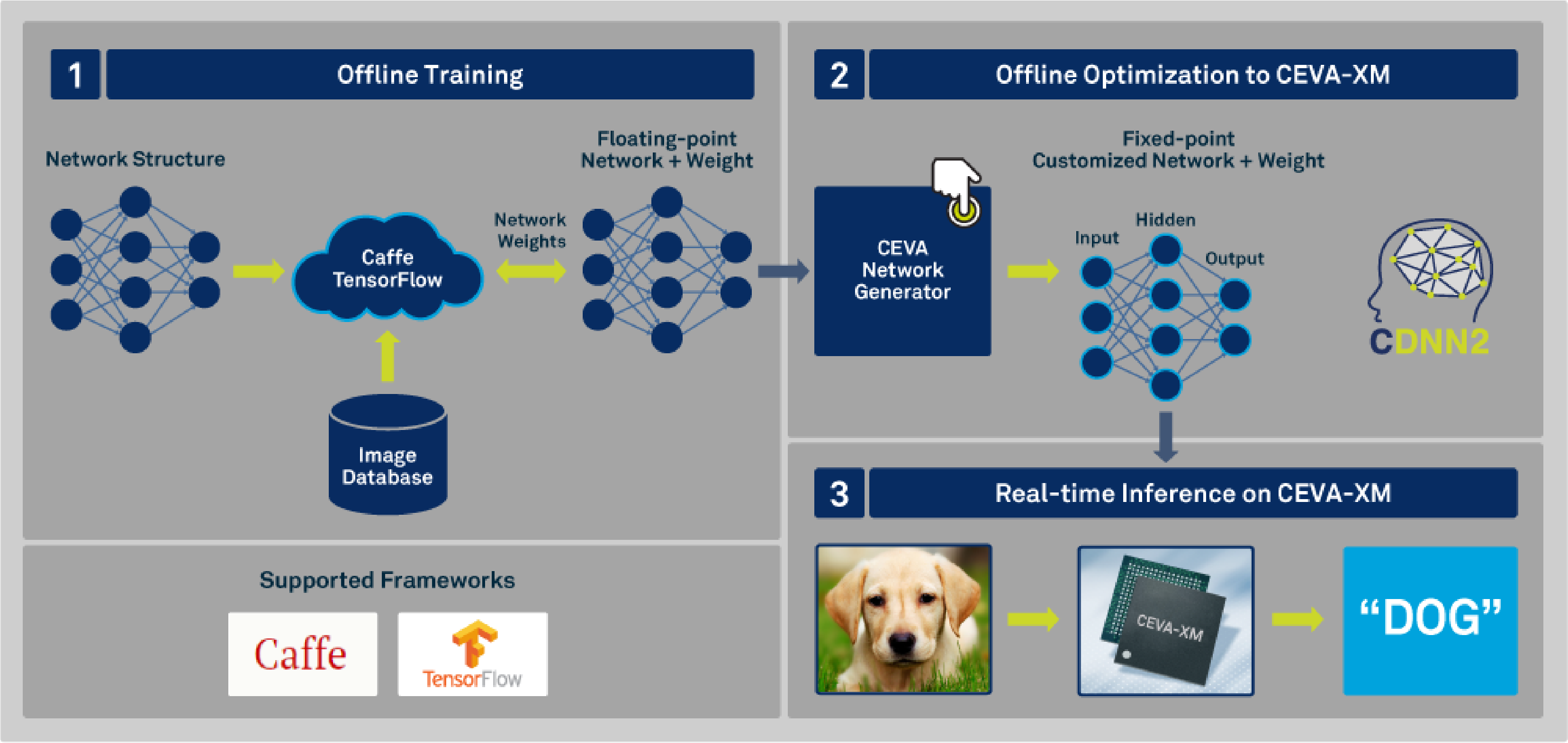
《中国电子商情》杂志社
2017年3月15日

March, 2017

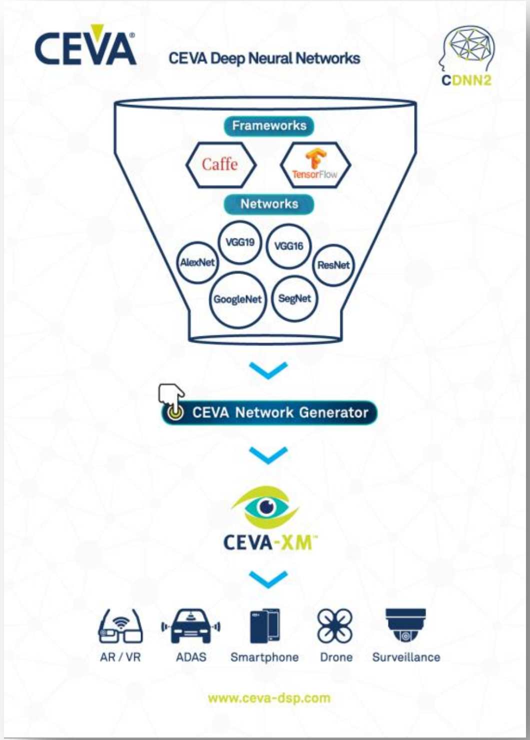
Outline

- Neural Network Introduction and Embedded Challenges
- CEVA Deep Neural Network (CDNN) Toolkit
- CDNN2 SW Framework
- CNN HWA
- CDNN Performance
- CDNN Roadmap

CDNN2 Usage Flow



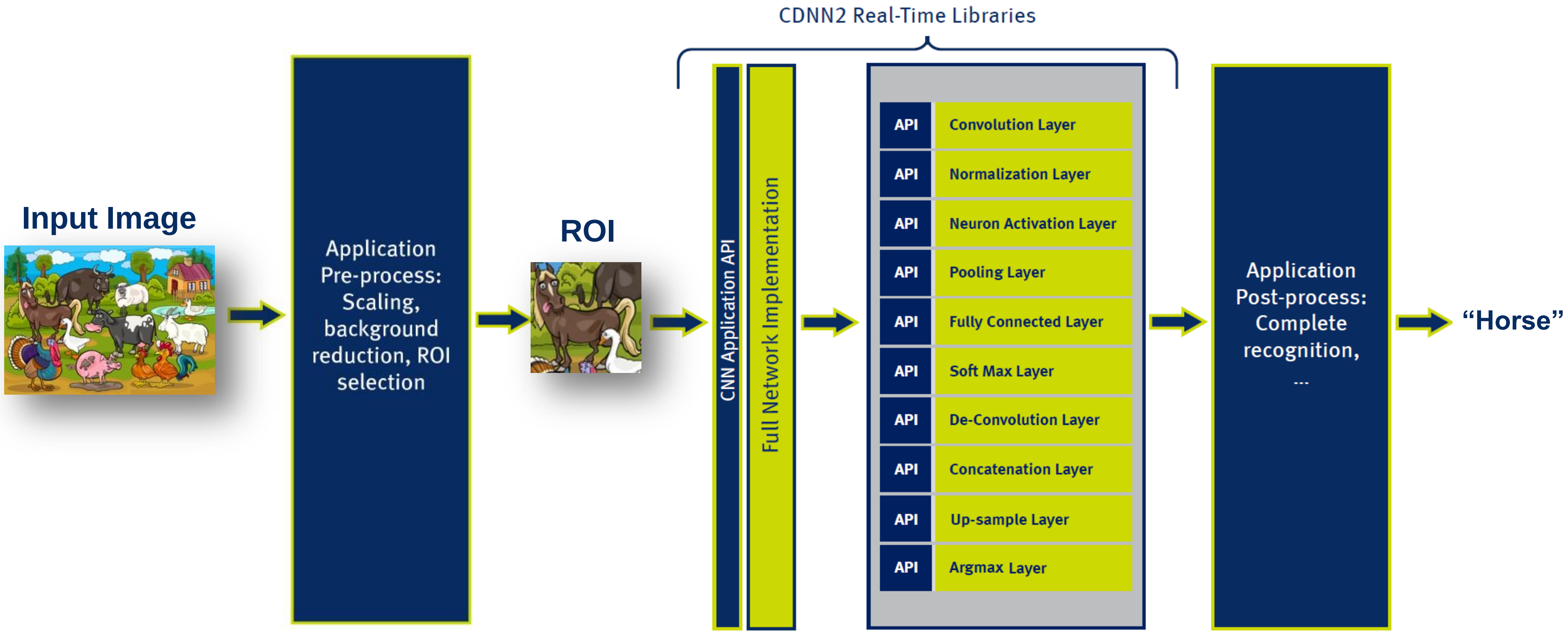
CEVA Deep Neural Network (CDNN2)



- 2nd gen SW framework support
  - Caffe and TensorFlow Frameworks
  - Various networks*
  - All network topologies
  - All the leading layers
  - Variable ROI
  - "Push-button" conversion from pre-trained networks to optimized real-time
  - Accelerates machine learning deployment for embedded systems
  - Optimized for CEVA-XM vision DSP together with CDNN HW accelerator

(*) Including AlexNet, GoogLeNet, ResNet, SegNet, VGG, NIN and others

Real-Time CDNN2 Application Flow



CDNN2 Feature Set

CEVA Network Generator (offline)	Neural Network Libraries (real-time)
Auto converts for power-efficiency	RT algo development and deployment
Floating to fixed point conversion	Optimized for CEVA-XM vision DSP
Adapts for embedded constraints	Various network structures and layers
Keeps high accuracy, 1% deviation	Fixed or variable input sizes
Caffe & TensorFlow support	On-the-fly bandwidth optimizations

Deliverables include real-time example models for image classification, localization, object detection

AlexNet Probabilities – Float vs. Fixed

CDNN2 CEVA®

| ▲ | <1%

Object	AlexNet PC Probability (floating point)	AlexNet on XM4 Probability (fixed point)
Labrador retriever	90.44%	**91.01%**
Golden retriever	4.45%	3.98%
Beagle	0.21%	**0.18%**
Kuvasz	0.12%	**0.10%**
Classification Probabilities		

See additional video comparing floating point to CDNN
https://www.youtube.com/watch?v=VnbCVFyuWYk

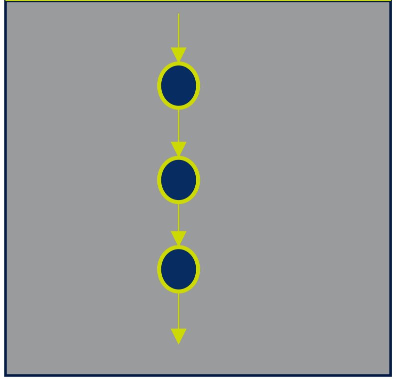
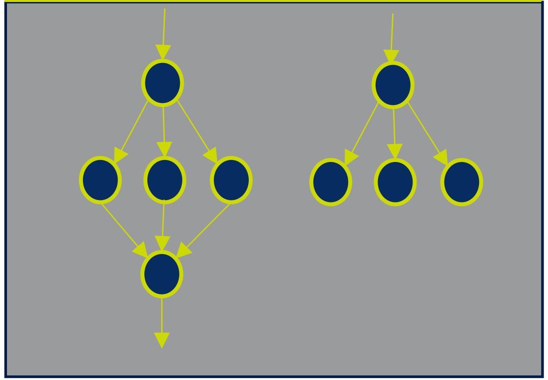
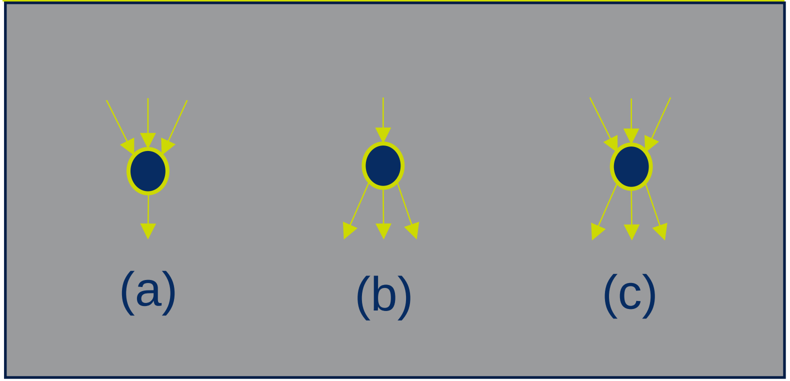
Caffe (32bit PC) Vs. CDNN2 (16bit Embedded)

https://youtu.be/VnbCVFyuWYk

CDNN2 Supported Topologies

CEVA CDNN2

- **All** network topologies are supported

	Linear Networks	Multiple Layers Per Level	Multiple-Input-Multiple-Output
Topology			
Networks	AlexNet VGG-19 VGG-16 VGG_S	GoogLeNet	GoogLeNet SegNet ResNet

CDNN2 Supported Networks

- CDNN2 supports the most advanced neural network including
  - Public Networks
    - Alexnet
    - CaffeNet
    - GoogleNet
    - ResNet
    - Yolo
    - Faster RCNN
    - Cifar10, Cifar10_nin
    - finetune_flickr_style
    - googlenet_finetune_web_car_iter_10000
    - googlenet_places205
    - KevinNet_CIFAR10_48
    - NIN
    - Pascal_VOC
    - VGG – 16,19, CNN_F, CNN_M, CNN_M_1024, CNN_M_128,CNN_M_20148, CNN_S, S
  - Proprietary Networks
    - From customers and partners under NDA

CDNN2 Supports over 80 advanced networks

CEVA-XM Advantages for Deep Learning

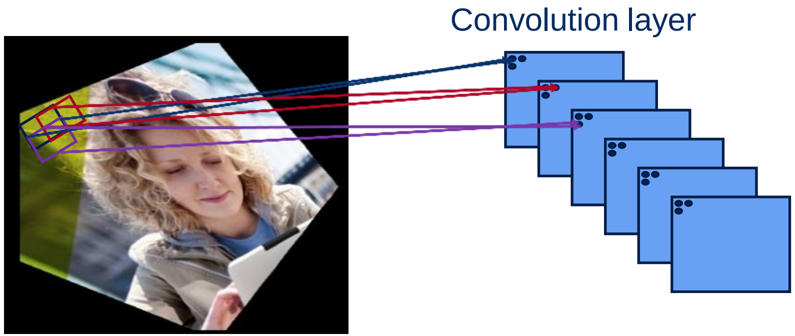
CEVA

Architectural Advantages

- CNN combines 2D convolutions, 2D max and 1D MAC operations
  - Efficient DSP can achieve great performance and power
- 2-Dimension data reuse fits 2D convolutions in CNN, enables high MACs/cycle utilization
- Neural Network entry point utilizes data reuse for lowering memory BW
- Parallel Random Memory Access – used for activation layer (Sigmoid, TanH)
- High precision accumulation required for fully connected layer

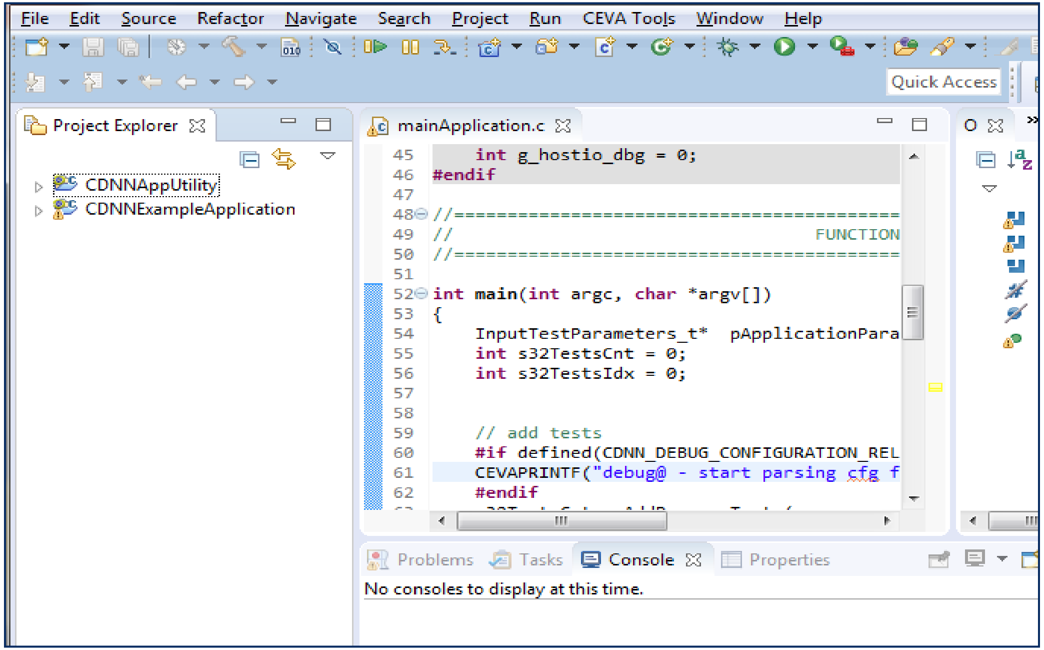
General Advantages

- CEVA-XM supplies flexible and scalable solution
  - Multi-cores scaling for higher requirements
  - Connectivity to additional accelerators (CEVA-Connect, AXI)
- Programmable solution ideal for evolving algorithms



CDNN2 PC Simulation Package

- Install CEVA-XM SDT CDNN Evaluation SW package
- Launch visual studio SW
- Import 2 example projects
  - There are 2 different projects, one is for windows and the other is for Linux
- Project → Build All to build the project
- Open pre defined ‘CDNN Debug Simulation’ debug configuration and push ‘Debug’ button to execute



Enable user getting neural network’s cycle count accuracy on PC without having a dedicated HW

CDNN – Developer Flow

Simplicity of running an application using CDNN

a. Create CDNN CEVA handle
   - CDNNCreate()

b. Create the network model (based on CDNN conversion tool outputs)
   - CDNNCreateNetwork()

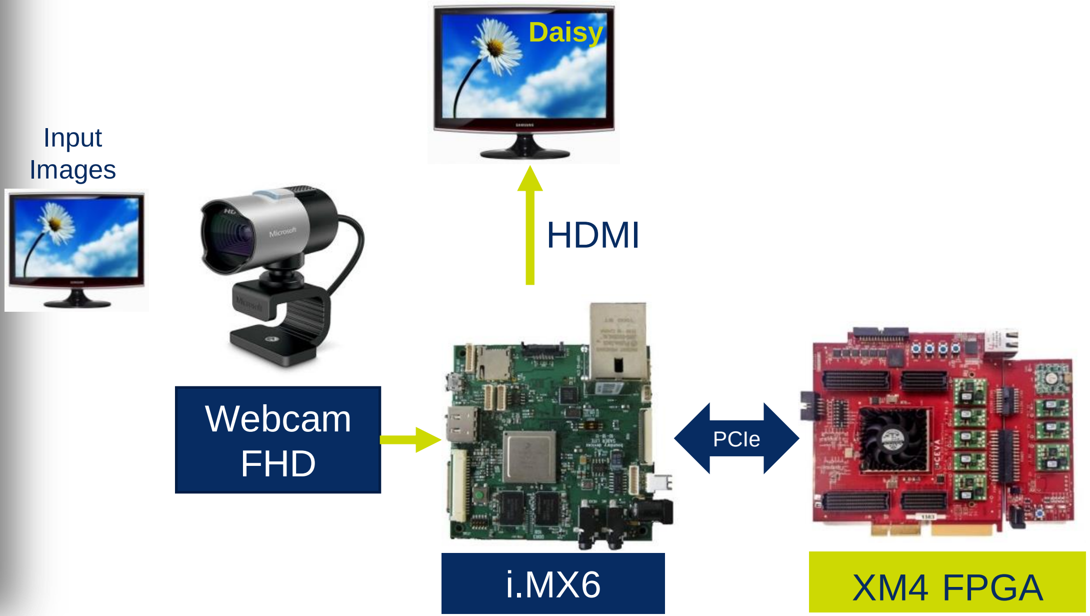
c. Initialize CDNN library (by creating a network and a memory database)
   - CDNNInitialize()

d. Execute the network (no need for re-initialization)
   - CDNNNetworkClassify()

Real-Time CNN Object Recognition Demo

[Live Alexnet object recognition]

- Enables <u>milli-watt</u> products vs. <u>watts</u> on GPU



CEVA Network Generator

Image Database
(For Training)

Caffe
TensorFlow

Pre-Trained Network
(*Deploy.caffemodel*)

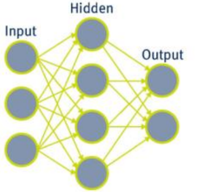
Network Structure
(Deploy.prototxt)

Normalization Image
(*Deploy.binaryproto*)

Network Labeling
(Deploy.labels)

CEVA
Network
Generator

CEVA-XM4
Optimized
Network



Live CDNN2 demo

Real-Time Network Generator Demo

CDNN2 CEVA

Live CDNN2 demo:
https://www.youtube.com/watch?v=SXINFryLM3Q&feature=youtu.be

Age and Gender Classification using Convolutional Neural Networks

Gil Levi Tal Hassner

The Open University of Israel

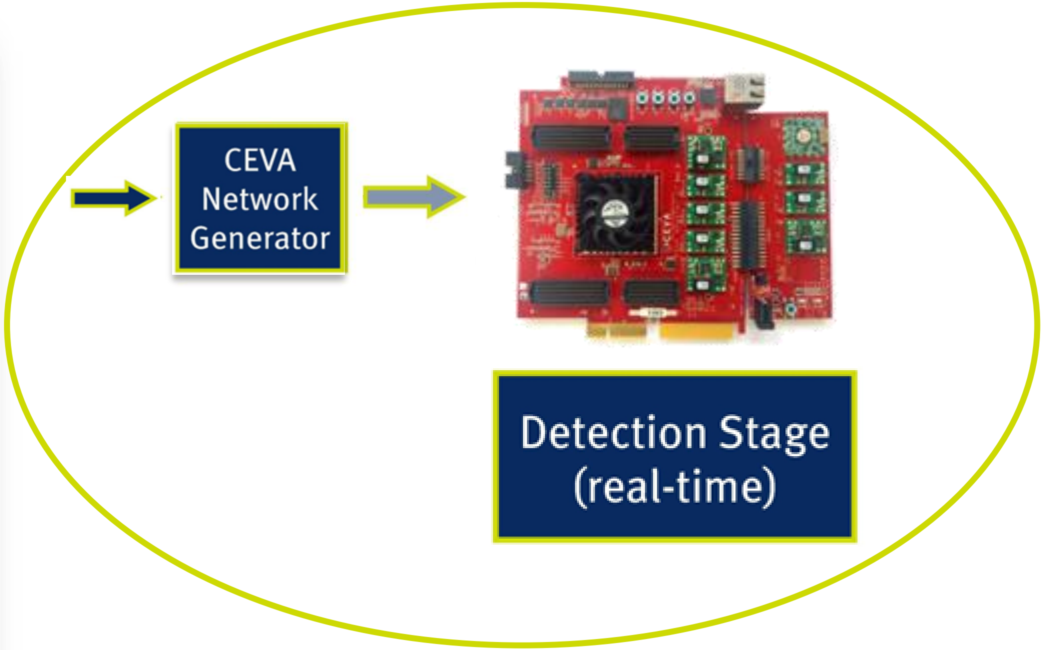
Figure 1. Faces from the Adience benchmark for age and gender classification. These images represent some of the challenges of age and gender estimation from real-world, unconstrained images. Most notably, extreme blur (low-resolution), occlusions, out-of-plane pose variations, expressions and more..

Abstract: Automatic age and gender classification has become relevant to an increasing amount of applications, particularly since the rise of social platforms and social media. Nevertheless, performance of existing methods on real-world images is still significantly lacking, especially when compared to the tremendous leaps in performance recently reported for the related task of face recognition. In this paper we show that by learning representations through the use of deep-convolutional neural networks (CNN), a significant increase in performance can be obtained on these tasks. To this end, we propose a simple convolutional net architecture that can be used even when the amount of learning data is limited. We evaluate our method on the recent Adience benchmark for age and gender estimation and show it to dramatically outperform current state-of-the-art methods..

Reference: Gil Levi and Tal Hassner, *Age and Gender Classification using Convolutional Neural Networks*, IEEE Workshop on Analysis and Modeling of Faces and Gestures (AMFG), at the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Boston, June 2015

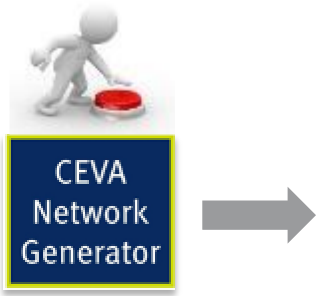
Click here for the PDF
Click here for the BibTex

Downloading Age classification
Neural Network from the internet



Passing it via CEVA Network Generator and running it on the XM4 FPGA under 10 min !

Example: AlexNet PC Profiler



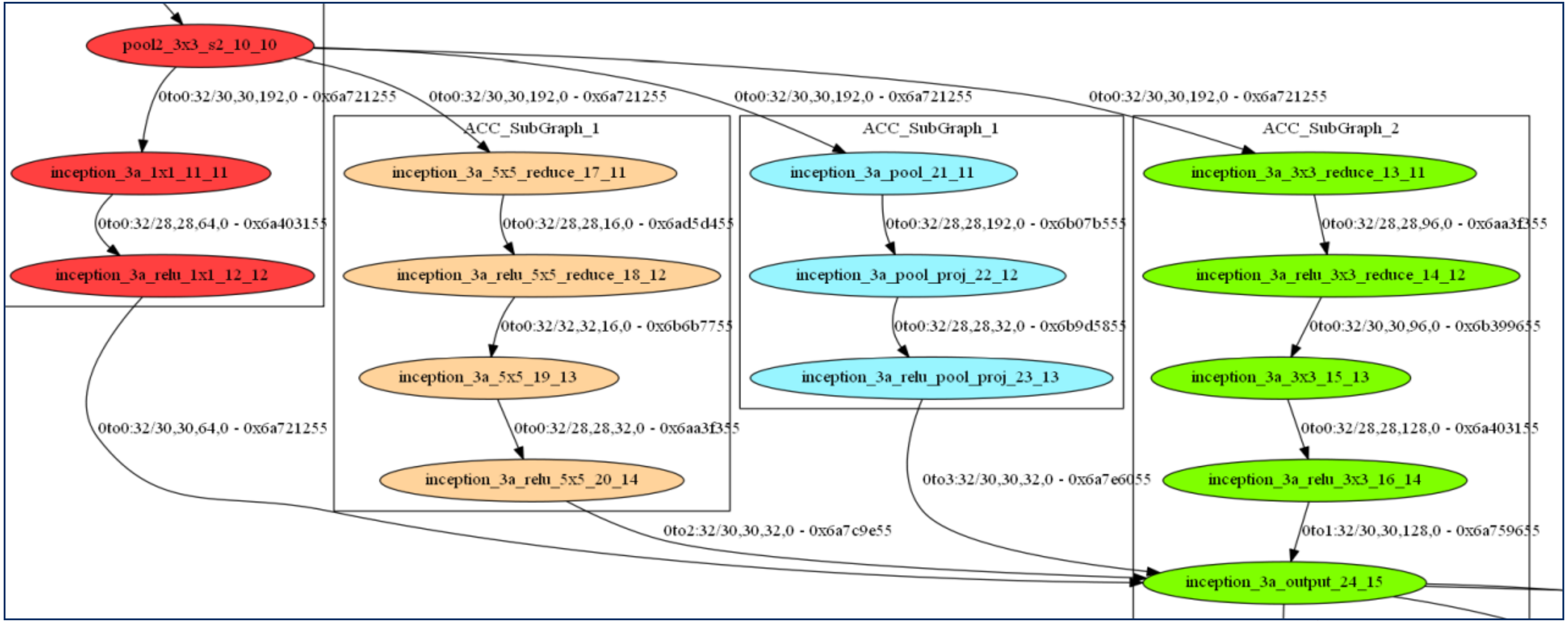
>>>> Network Structure

Layer ID:	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Layer Type:	InputLayer	ConvLayer	NormLayer	PoolLayer	ConvLayer	NormLayer	PoolLayer	ConvLayer	ConvLayer	ConvLayer	PoolLayer	FullyConnectedLayer	FullyConnectedLayer	FullyConnectedLayer	CrossChannelOperationLayer
Layer Name:	input	conv1	norm1	pool1	conv2	norm2	pool2	conv3	conv4	conv5	pool5	fc6	fc7	fc8	prob
Input Number:	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Input Dimension X:	612	227	55	55	27	27	27	13	13	13	13	6	1	1	1
Input Dimension Y:	612	227	55	55	27	27	27	13	13	13	13	6	1	1	1
Num. of input maps:	3	3	96	96	96	256	256	256	384	384	256	256	4096	4096	1000
Kernel Dimension X:	0	11	5	3	5	5	3	3	3	3	3	0	0	0	0
Kernel Dimension Y:	0	11	5	3	5	5	3	3	3	3	3	0	0	0	0
Padding Dimension X:	0	0	0	0	2	0	0	1	1	1	0	0	0	0	0
Padding Dimension Y:	0	0	0	0	2	0	0	1	1	1	0	0	0	0	0
Stride Dimension X:	0	4	0	2	1	0	2	1	1	1	2	0	0	0	0
Stride Dimension Y:	0	4	0	2	1	0	2	1	1	1	2	0	0	0	0
Output Number:	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Output Dimension X:	227	55	55	27	27	27	13	13	13	13	6	1	1	1	1
Output Dimension Y:	227	55	55	27	27	27	13	13	13	13	6	1	1	1	1
Num. of output maps:	3	96	96	96	256	256	256	384	384	256	256	4096	4096	1000	1000
Pooling Mode:				max			max				max				
Activation Mode:		Relu			Relu			Relu	Relu	Relu		Relu	Relu		
K:	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0
Alpha:	0	0	0.0001	0	0	0.0001	0	0	0	0	0	0	0	0	0
Beta:	0	0	0.75	0	0	0.75	0	0	0	0	0	0	0	0	0
Dropout Factor:	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

>>>> Network Statistics

BW Reduction	0.1														
NumberOfInputChannels	3	3	96	96	96	256	256	256	384	384	256	256	4096	4096	1000
NumberOfInputZeroChannels	0	0	0	0	0	4	4	4	1	1	13	121	3545	3857	0
NumberOfInputNonZeroElements	1123619	154576	143937	143736	63452	35572	35541	20048	21327	20096	3856	706	551	239	1000
NumberOfLayerWeights	0	35712	0	0	491520	0	0	1179648	884736	589824	0	37748736	16777216	4096000	0
NumberOfBytesPerWeight	0	2	0	0	2	0	0	2	2	2	0	1	1	1	0
NumberOfLoadedWeights	0	35712	0	0	491520	0	0	1179648	884736	589824	0	2891776	2256896	239000	0
Weights BW	0	71424	0	0	983040	0	0	2359296	1769472	1179648	0	2891776	2256896	239000	0
Total Weight BW	11750552														
Internal memory size[B]	524288														
Input memory type internal/external	External	External	External	External	Internal	External	External	Internal	External	External	External	Internal	Internal	Internal	Internal
NumberOfInputElements	1123632	154587	290400	290400	92256	186624	186624	57600	86400	86400	43264	9216	4096	4096	1000
NumberOfBytesPerElement	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Input BW	1123632	309174	580800	580800	0	373248	373248	0	172800	172800	86528	0	0	0	0
Total Input BW	3773030														
Output memory type internal/external	External	External	External	Internal	External	External	Internal	External	External	External	Internal	Internal	Internal	Internal	Internal
NumberOfOutputElements	154587	290400	290400	92256	186624	186624	57600	86400	86400	43264	9216	4096	4096	1000	1000
NumberOfBytesPerElement	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2
Output BW	309174	580800	580800	0	373248	373248	0	172800	172800	86528	0	0	0	0	0
Total Output BW	2649398														
Total input/output BW	6422428														
Total input/output/weights BW	18172980														

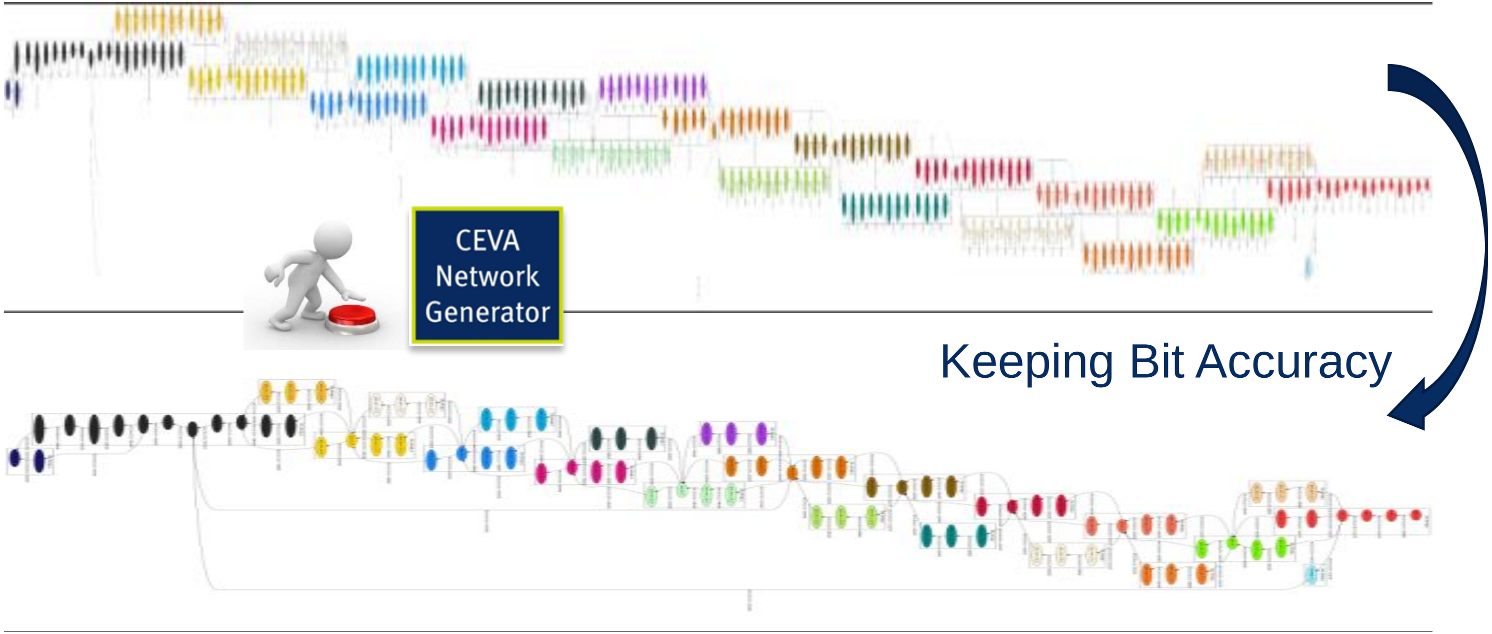
Example: GoogleNet Challenge



Example: FasterRCNN Challenge

Full automatic network analysis and optimization without any user involvement

Before
High BW



Keeping Bit Accuracy

After
Low BW

Outline

- Neural Network Introduction and Embedded Challenges
- CEVA Deep Neural Network (CDNN) Toolkit
- CDNN2 SW Framework
- CNN HWA
- CDNN Performance
- CDNN Roadmap

CEVA-CNN HW Accelerator

- **Motivation**
  - Convolutions are the major and most cycles consuming layers
  - Dedicated HW engine for executing the **convolutions** layers in CNN
  - Provides the flexibility to cope with future Neural Network development

>8X

Performance Gain*

- **Compatibility**: CEVA-XM vision processors

* vs. CEVA-XM4

Flexible Embedded CNN Solution

CEVA

CEVA-XM Vision DSP

CDNN2 Real-Time SW Library

- Controls Full network execution
- Invoke CNN HWA
- Executes all other layers:
  Normalization,
  Pooling,
  Deconvolution,
  Etc.
- Supports Multiple CNN HWAs

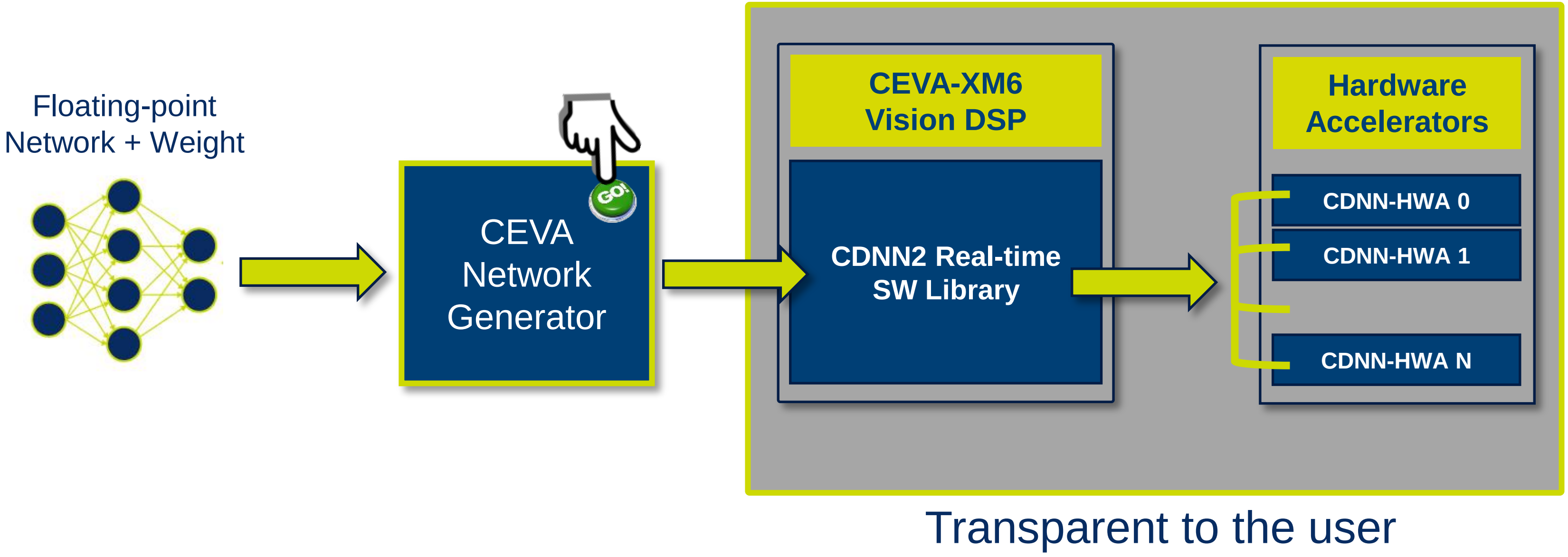
CNN Hardware Accelerator

CNN HWA V1

- Up to 520 MACs units
  (130/260/520 MACs)
- 16b x 16b Support
- Executes Convolutions
- Internal Memories
- Internal DMA units
- Autonomous execution

Flexible embedded solution and 16bit support are required to cope with the evolving and leading neural networks

Automatic Usage of Multiple HWAs



CNN HWA Schedule

- RTL
  - Beta version by Feb 2017
  - Final version by April 2017
- SW Support (CDNN2 V3.0.0.F)
  - XM4 and XM6 – June 2017

Outline

- Neural Network Introduction and Embedded Challenges
- CEVA Deep Neural Network (CDNN) Toolkit
- CDNN2 SW Framework
- CDNN HWA
- CDNN Performance
- CDNN Roadmap

CDNN2 Performance

CEVA

Core	L1 Data Size	AlexNet Perf (1000 classes, 227 x 227)			Tiny YOLO (16x16b , 448 x 448)				Small YOLO (16x16b , 448 x 448)			
		MC/ Image	BW / Image (MB)	ROI/SEC @600MHz	MC/ Image	BW / Image (MB)	Ext. memory (MB)	ROI/SEC @600MHz	MC/ Image	BW / Image (MB)	Ext. memory (MB)	ROI/SEC @600MHz
XM4	512KB	20	18	30	75	425	82.7	8	580	4GB	82.9	1
XM6	512KB	11.5	18	52	38	425	82.7	15	290	4GB	82.9	2
XM4 + 520 HWA	256KB + 1152KB HWA	4.8 core 1.3 HWA	11 core 5.6 HWA	125	5 core 5.5 HWA	38 core 68 HWA	82.7	109	17.1 core 45 HWA	64 core 199 HWA	82.9	13
XM6 + 520 HWA	256KB + 1152KB HWA	3.4 core 1.3 HWA	11 core 5.6 HWA	176	4.2 core 5.5 HWA	38 core 68 HWA	82.7	109	12.2 core 45 HWA	64 core 199 HWA	82.9	13
XM4 + 520 HWA	256KB + 2MB HWA	4.8 core 1.3 HWA	11 core 5.6 HWA	125	5 core 5.5 HWA	38 core 67.8 HWA	82.7	109	17.1 core 45 HWA	64 core 172 HWA	82.9	13
XM6 + 520 HWA	256KB + 2MB HWA	3.4 core 1.3 HWA	11 core 5.6 HWA	176	4.2 core 5.5 HWA	38 core 67.8 HWA	82.7	109	12.2 core 45 HWA	64 core 172 HWA	82.9	13

CEVA-XM6 Platform vs. NVidia TX1 GPU for Implementing Deep Learning

- Single CEVA-XM6 based platform is

Power Efficiency Factor*

>25X

>4X

Faster Processing**

Assumptions:

- Based on the implementations of AlexNet and GoogleNet (single batch)
- TSMC 20nm technology and core @690MHz
- (*) ROI/Sec/Watt (**) ROI/Sec
- Nvidia TX1 information: https://www.nvidia.com/content/tegra/embedded-systems/pdf/jetson_tx1_whitepaper.pdf

Outline

- Neural Network Introduction and Embedded Challenges
- CEVA Deep Neural Network (CDNN) Toolkit
- CDNN2 SW Framework
- CDNN HWA
- CNN Performance
- CDNN Roadmap

CEVA CNN Roadmap

CEVA

Release Version	Target Date
CEVA-XM4 CDNN2 v2.2.1 - Repack with license	Available
CEVA-XM6 CDNN2 v2.2.2 - XM6 Support	Available
CNN HWA RTL v1.0.0	Available
CEVA-XM4 CDNN2 v3.0.0 – see separate slide CEVA-XM6 CDNN2 v3.0.0 – see separate slide	Jun 20th,2017
CEVA-XM6 CDNN2 v3.0.1 – XM6 Optimized	Aug 31th,2017
CEVA-XM4 CDNN2 v4.0.0 – see separate slide CEVA-XM6 CDNN2 v4.0.0 – see separate slide	Dec 31th,2017

CEVA-XM CDNN2 v3.0.0 – June 20th,2017

CEVA®

- Integration with CNN HWA
- Enhanced TensorFlow support
- Real-time Dynamic Precision
- Faster RCNN Optimized

XM CDNN2 v4.0.0 – December 2017

- Weights compression
- 8 bit networks
- Additional layers support
- RNN
- Custom Layer Support
- Multicore support

CEVA-XM CDNN Toolkit Summary

CEVA

Key Differentiation
Comprehensive Solution Best balanced solution between HWA, DSP and SW to allow most efficient and progressive solution in terms of area, performance efficiency and short time to market

SW Support
• CDNN SW framework allows short "time-to market"
    • CEVA Network Generator – 2nd generation
    • CDNN2 real-time library – 2nd generation

Configurable Solution • 130/260/520 16x16b MACs units options
Flexible and Optimized Solution • Support variable kernel sizes and input dimensions • New layers can be added and executed easily on the XM • Compression/decompression technique are on the roadmap as well as many others improvements • CNN HWA is working directly with the memory ➔ no need for additional accumulators / resources and no impact on the utilization
Maturity and Availability • Supports and runs the most advanced NNs layers and networks • Available today

The DSP Powerhouse

Thank You

www.ceva-dsp.com

Resources

- The Ultimate Deep Learning & Artificial Intelligence Platform for Low-power Embedded Devices
- CEVA Deep Neural Network (CDNN) product page
- CEVA CDNN live AlexNet demonstration
- CEVA CDNN2 Network Generator live demonstration
- Automotive “Free Space” using CDNN2 demonstration
- Caffe (32bit PC) Vs. CDNN2 (16bit Embedded)
- CDNN2 Webinar