



CEVA[®]



CEVA Deep Neural Network

June 2016

www.ceva-dsp.com

CEVA Deep Neural Network (CDNN2)



- ▶ 2nd gen SW framework

- ▶ Accelerates machine learning deployment for embedded systems
- ▶ Optimized for CEVA-XM4 vision DSP

- ▶ Most popular deep learning frameworks supported:

1. **Google's TensorFlow** based networks

2. **Caffe** based networks

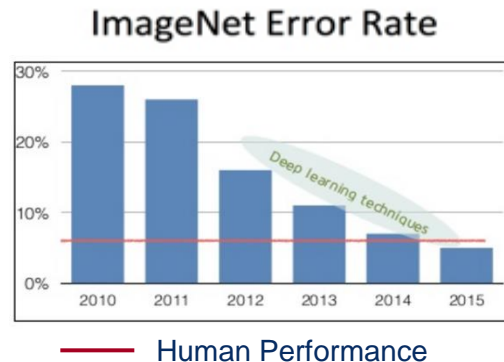
- ▶ Variety of layers – including Deconvolution, Concatenation, Up-sample & Argmax
- ▶ Various topologies – Linear, multiple layers per level, Multiple-Input-Multiple-Output and fully convolutional networks
- ▶ Various networks – including AlexNet, GoogLeNet, ResNet, SegNet, VGG, NIN

- ▶ Push-button conversion from pre-trained networks to optimized real-time

Deep Learning Performance Improvements



- ▶ Until recently researchers were limited mainly by computing horsepower, power constraints and algorithmic quality
- ▶ Big progress on these fronts recently
- ▶ Deep learning has achieved the state-of-the-art in areas like image classification, speech, and natural language processing
- ▶ ImageNet - a dramatic 4x improvement over the past five years.
 - ▶ Deep learning techniques achieved a 16% top-5 error rate in 2012 and are now below 5%, **exceeding human performance**



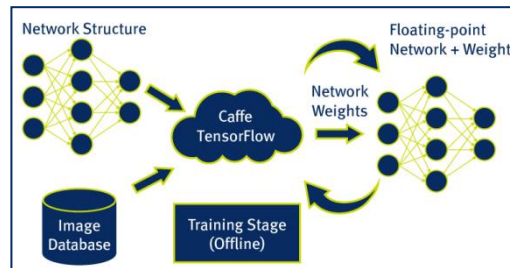
Source: [Nervana Systems](#)

Leading Deep Learning Frameworks



Caffe

- ▶ Is a well-known and widely used machine-vision library that ported Matlab's implementation of fast convolutional nets to C and C++
- ▶ Made with expression, speed, and modularity in mind
- ▶ Used by both researchers, academy, and some commercial use



- ▶ Relatively new alternative to Caffe supported and promoted by Google
- ▶ Scalable to work both for research and commercial purpose without making any changes
- ▶ A software library for numerical computation using data flow graphs

Caffe vs. TensorFlow (TF)

► Maturity

- The Caffe framework benefits from having a large repository of pre-trained neural network models suited for a variety of image classification tasks, called the **Model Zoo**. TF was first introduced during Nov 2015

► Applicability

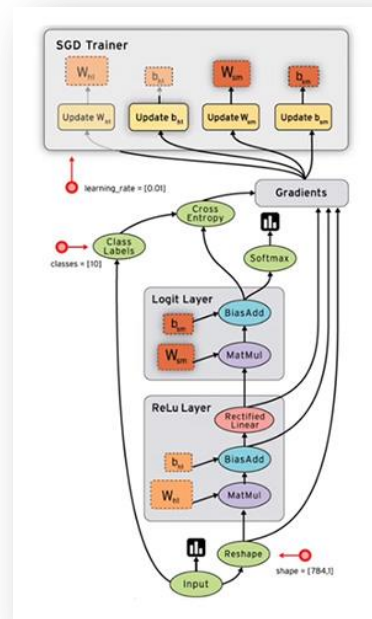
- Caffe is used for image classification and not intended for other deep-learning applications such as text or sound, while TF addresses general applications

► Modeling Capabilities

- Caffe support for recurrent networks (RNN) and language modeling is generally poor, due to its legacy architecture, while RNN API and implementation are suboptimal in TF
- Caffe is not considered flexible because for new layer types you have to define the full forward, backward, and gradient update
- Since TF uses symbolic graph of vector operations approach, specifying a new network is fairly easy

► Architecture

- TF has a cleaner, modular architecture with multiple frontends and execution platforms

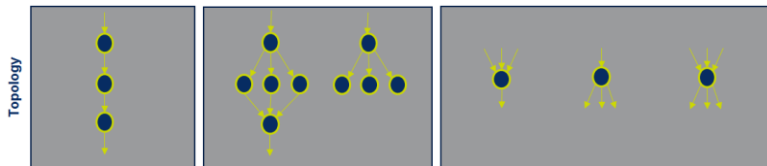


The Deep Neural Networks Challenge

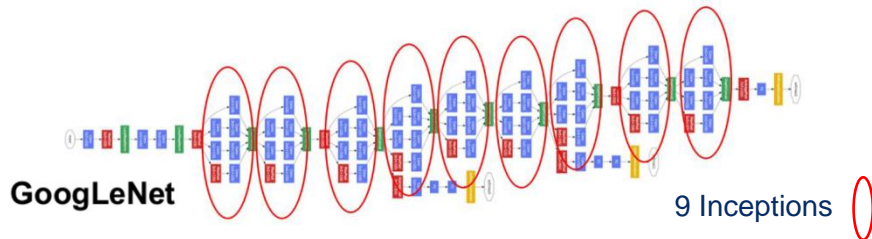


CEVA®

- ▶ Implementing a deep neural network in embedded systems is a challenging task
- ▶ Typical deep neural network could exhaust up to Giga bytes of memory and result in bandwidth and computing bottleneck
- ▶ Deep neural networks include different training frameworks, different layers and different network topologies



- ▶ Need to deal with Network-in-network (NIN)



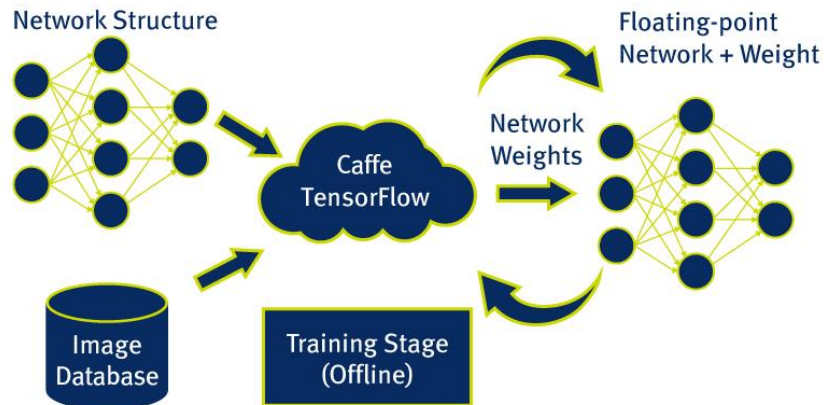
CEVA Deep Neural Network (CDNN2)



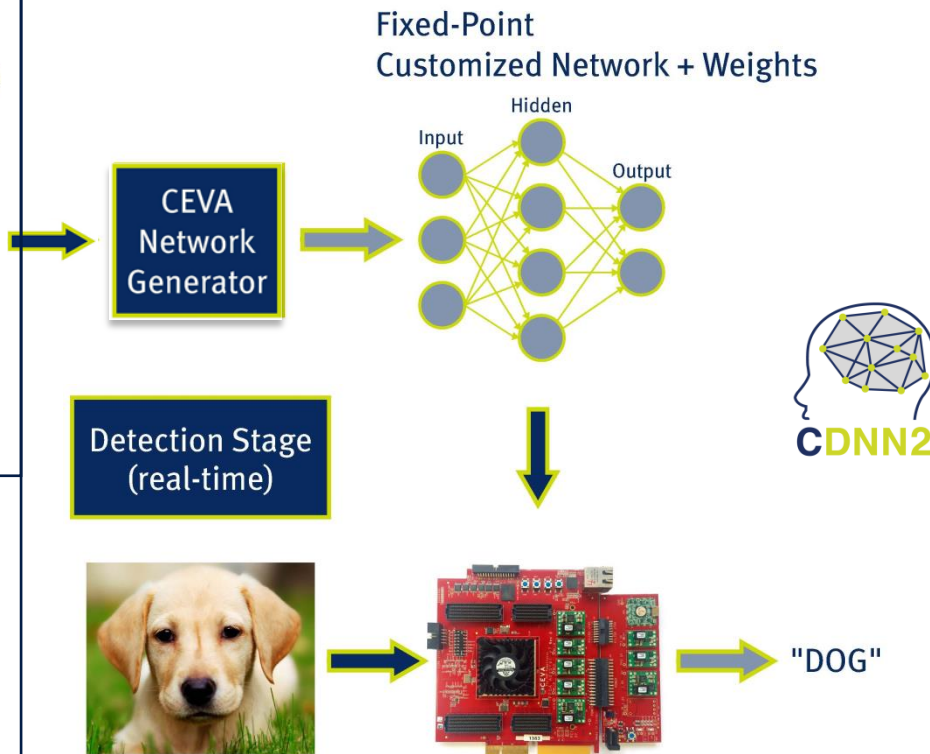
- ▶ A software framework providing real-time, efficient object recognition & vision analytics
 - ▶ Harnessing the power of the CEVA-**XM4** imaging & vision DSP
 - ▶ Lowest power and memory bandwidth deep learning solution
 - ▶ Significant time-to-market advantage for implementing NN in embedded systems (see separate video clip)
- ▶ Enables real-time classification with pre-trained networks
 - ▶ Receives pre-trained network model & weights as input (via “Caffe” or “TensorFlow”)
 - ▶ Automatically converted into a real-time network model, via **CEVA Network Generator**
 - ▶ Utilizes real-time network model in CNN applications on CEVA-**XM4**

CDNN2 Usage Flow

OEM / Partner (offline)



CEVA (offline + real-time)



CDNN2 Feature Set



CEVA Network Generator (offline)

- ▶ Auto converts for power-efficiency
- ▶ Floating to fixed point conversion
- ▶ Adapts for embedded constraints
- ▶ Keeps high accuracy, 1% deviation
- ▶ Caffe & TensorFlow support

Neural Network Libraries (real-time)

- ▶ RT algo development and deployment
- ▶ Optimized for CEVA-**XM4** vision DSP
- ▶ Various network structures and layers
- ▶ Fixed or variable input sizes
- ▶ On-the-fly bandwidth optimizations

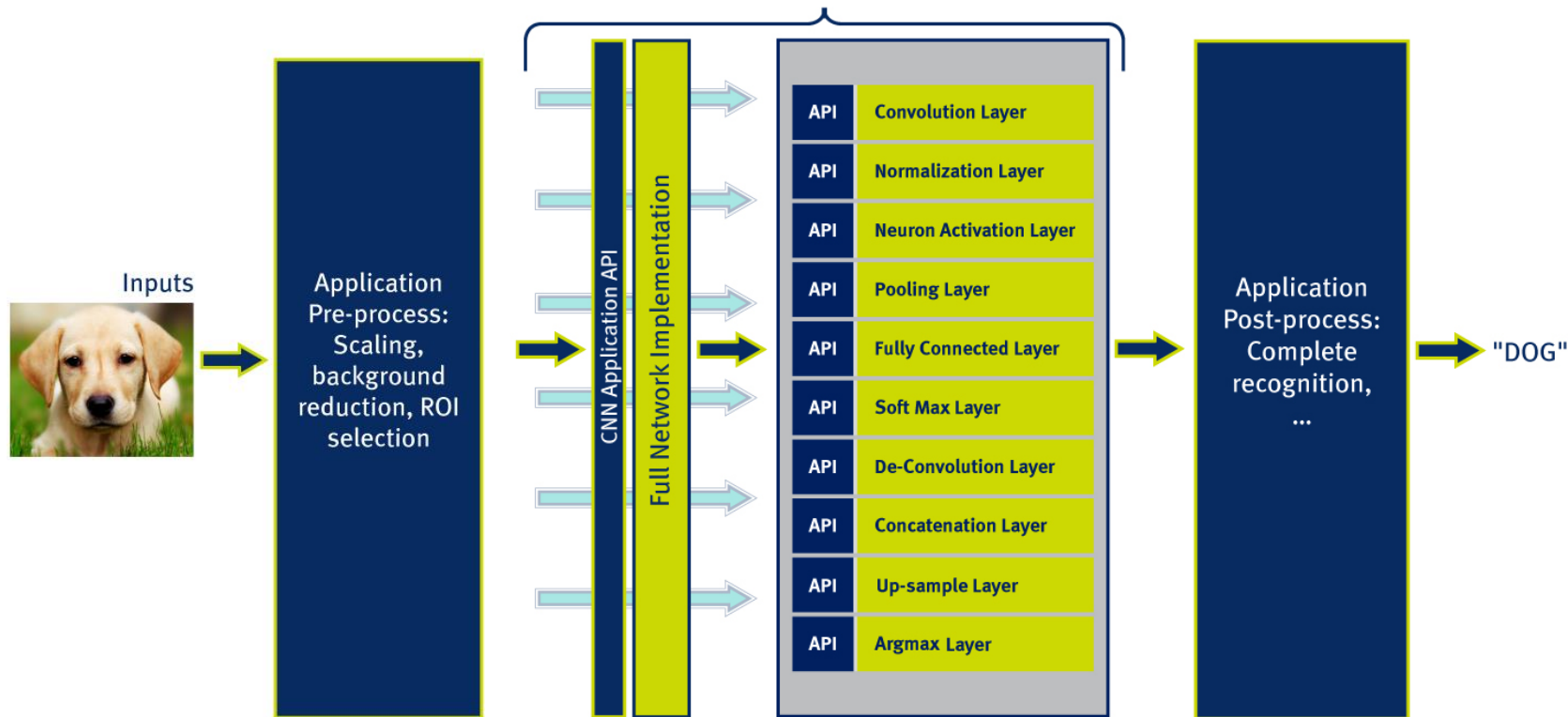
Deliverables include real-time example models for image classification, localization, object detection

Real-Time CDNN2 Application Flow



CEVA®

CDNN2 Real-Time Libraries



CDNN2 Supported Layers

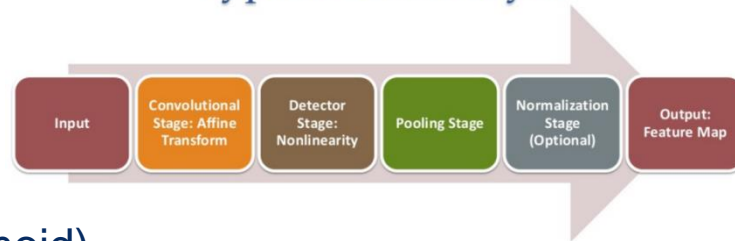


CEVA®

- ▶ CDNN2 supports the most advanced neural network layers including
 - ▶ Input manipulation layer (pre-process stage resize, jittering and more)
 - ▶ Convolutional
 - ▶ Normalization
 - ▶ Pooling (Average and Max)
 - ▶ Fully Connected
 - ▶ Softmax
 - ▶ Activation (ReLU, Parametric ReLU, TanH, Sigmoid)
 - ▶ **Deconvolution**
 - ▶ **Concatenation**
 - ▶ **Upsample**
 - ▶ **Argmax**
 - ▶ Custom user layer, attaching a specific functionality



Typical CNN Layer


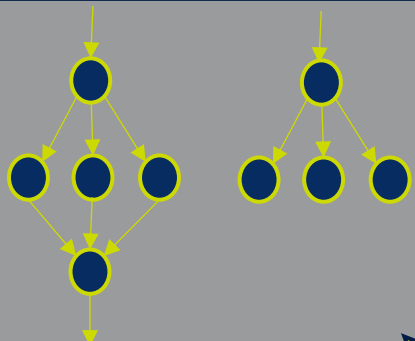





CDNN2 Supported Topologies



CEVA[®]

► **All** network topologies are supported

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

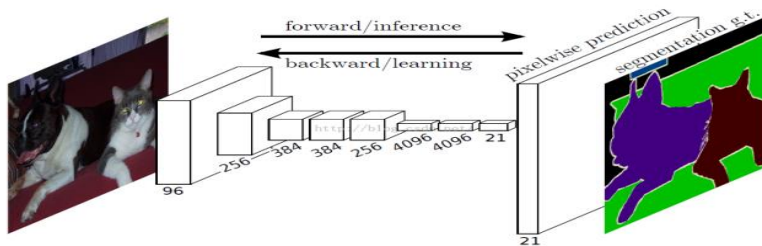
New

Fully Convolutional Networks



CEVA[®]

- ▶ CDNN2 supports Fully Convolutional Networks (FCNs)
- ▶ Fully convolutional networks are fast, end to-end models for pixel wise problems
- ▶ FCN can take input of arbitrary size and produce correspondingly-sized output with efficient inference and learning. This is more suitable for commercial applications where the ROI (Region of Interest) is dynamically changing depending on the object size. Saves cycles by adapting CNN to actual image size
- ▶ Example: AlexNet, VGG net, and GoogLeNet



CDNN2 Supported Networks*



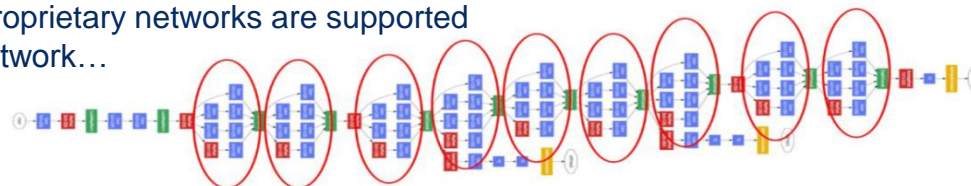
CEVA®

Network	# Layers	ROI	Special Layers	Topology
AlexNet	24	224x224		Linear
SegNet	90	480x360	Upsample, Argmax, Pooling with index	Multiple-Input-Multiple-Output
GoogLeNet	23 layers + 9 Inception**	220x220	Concatenation	Multiple-Input (concatenation layer) Multiple layers per level
VGG-19	19	224x224		Linear
VGG-16	16	224x224		Linear
VGG_S	24	224x224		Linear



- (*) 1. All Caffe and TensorFlow networks based are supported
 2. The above networks are running on FPGA or planned to by run until end of July
 3. The above list is partial, additional proprietary networks are supported
 (**) Inception: Network in a network in a network...

GoogLeNet



Convolution
 Pooling
 Softmax
 Other
 Inceptions

AlexNet - Network Performance



CEVA®

► Network specification

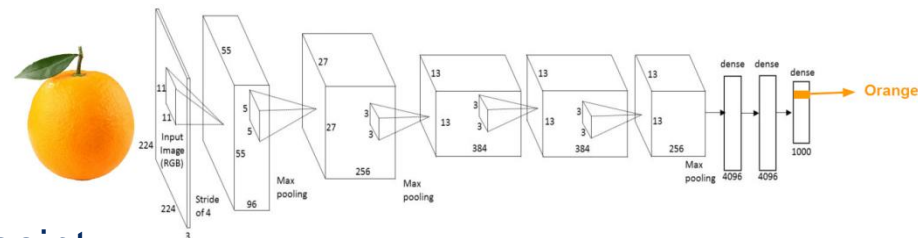
- Full forward classification case image measurement (single iteration)
- 24 layers, 224x224 network input size
- 11x11, 5x5 and 3x3 convolution filters

► Memory bandwidth

- Pre-trained network: 253Mbytes floating point
- Post CDNN2 (optimized for CEVA-XM4): 16Mbytes fixed-point
- Including weights and data

► Performance

- CEVA-XM4: 24 Mcycles
- CEVA-XM6: 12 Mcycles (Conv 94.7%, FC 4%, pool 0.3%, other 1%)



GoogLeNet - Network Performance



CEVA®

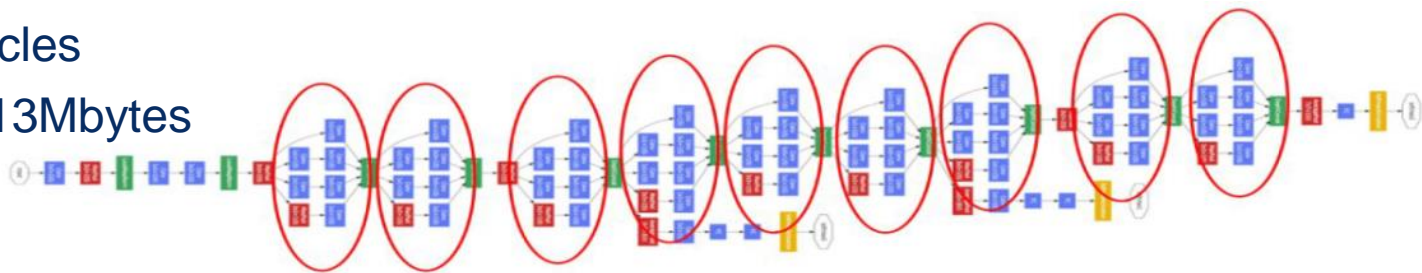
► Specification

- 32 Layers
- 224x224 network input size
- 7x7, 5x5, 3x3 and 1x1 convolution filters
 - XM4: 0.8 cycles per filter of 7x7
 - XM6: 0.4 cycles per filter

► Performance estimation for single execution:

- XM4: 39Mcycles
- XM6: 20Mcycles
- Weight BW ~13Mbytes

Convolution
Pooling
Softmax
Other
Inceptions



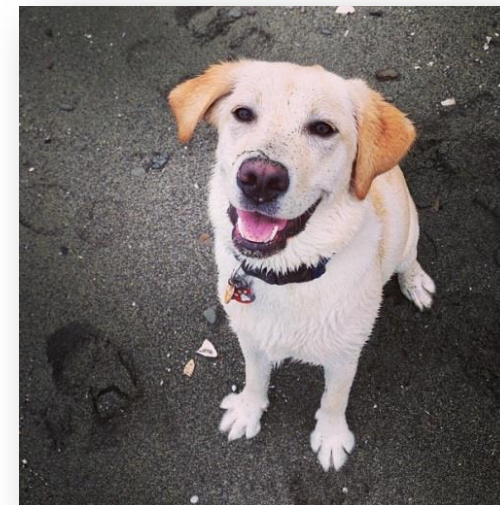
AlexNet Probabilities – Float vs. Fixed



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		



CEVA-XM4 Advantages for Deep Learning

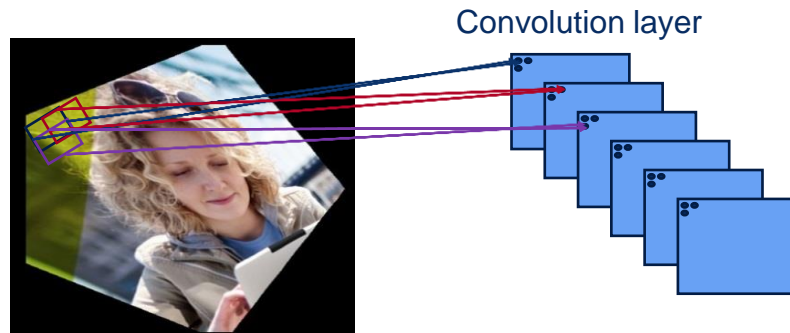


Architectural Advantages

- ▶ 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 utilization of 128MACs/cycle
- ▶ 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

- ▶ XM4 supplies flexible and scalable solution
 - ▶ Multi-cores scaling for higher requirements
 - ▶ Connectivity to additional accelerators (CEVA-Connect, AXI)
- ▶ Programmable solution ideal for evolving algorithms



Real-Time Network Generator Demo



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

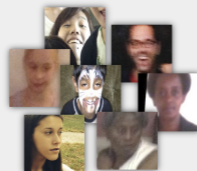


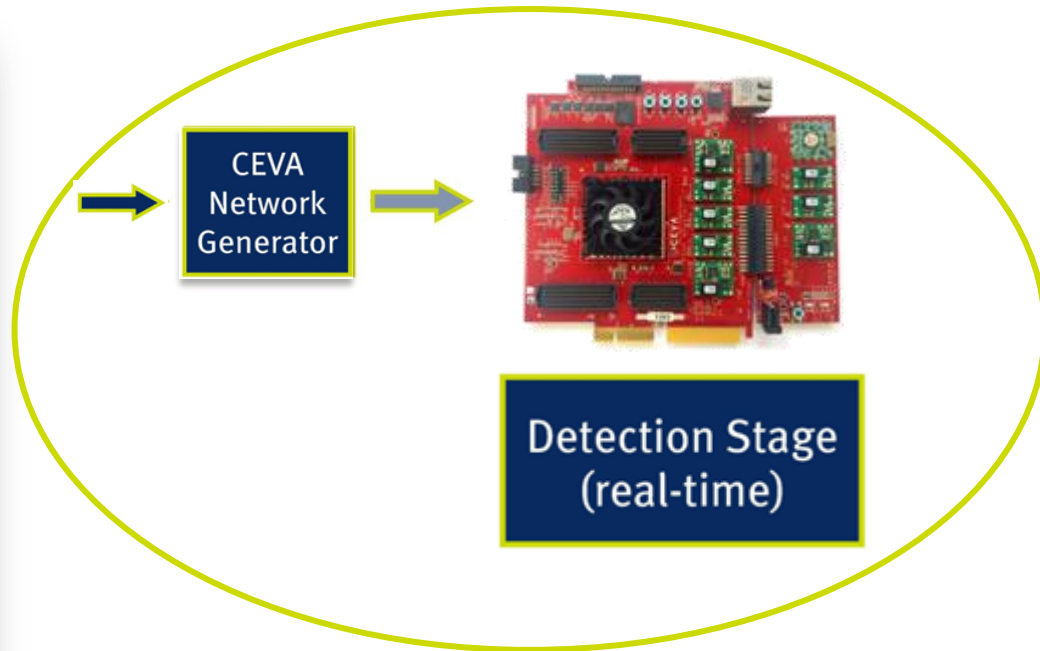
Figure 1. Faces from the [Adience benchmark](#) for age and gender classification. These images represent some of the challenges of age and gender estimation from real-world, unconstrained images. Most notably, extreme blur (low-resolution), occlusions, out-of-plane pose variations, expressions and more.

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](#)

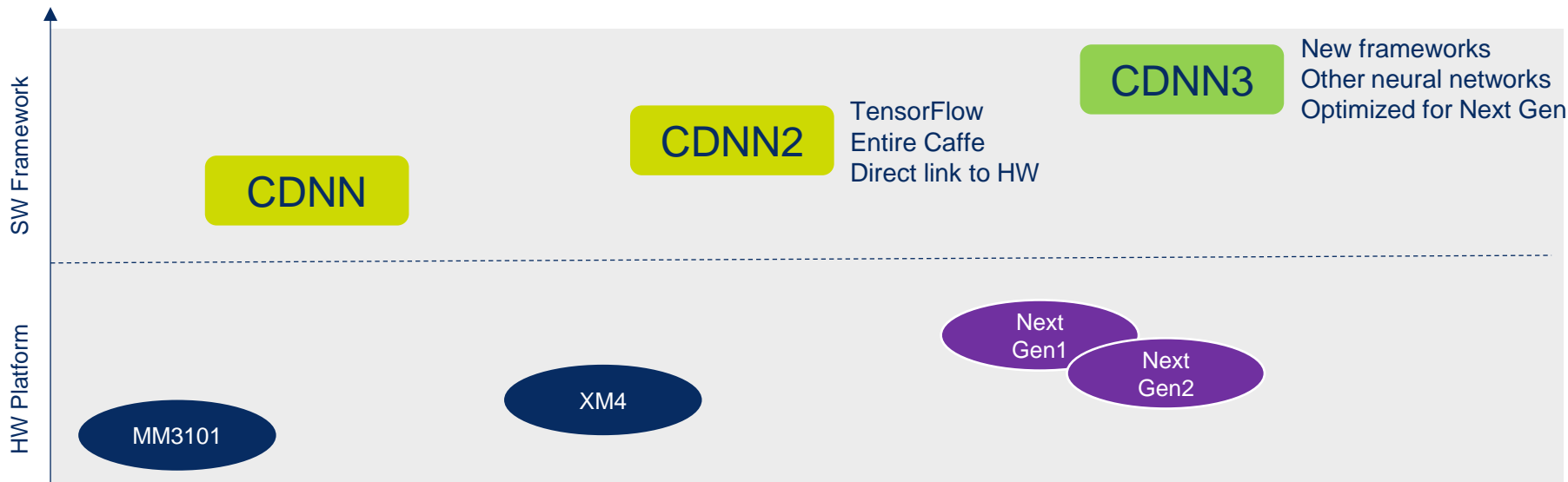
Downloading Age classification
Neural Network from the internet



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

Final Comments: HW + SW

- ▶ CEVA is at the forefront of development of Neural Network embedded platforms
- ▶ Usually the HW platform is meaningless if not supported by the corresponding SW framework...





Thank You

www.ceva-dsp.com