Object recognition and computer vision using MATLAB and NVIDIA Deep Learning SDK

17 May 2016, Melbourne
24 May 2016, Sydney

Werner Scholz, CTO and Head of R&D, XENON Systems
Mike Wang, Solutions Architect, NVIDIA
Outline

- Background of XENON System
- Object recognition - MATLAB interprets image contents
- How does it do it? Looking under the hood
- What is required? Software stack and installation
- NVIDIA Deep Learning SDK
- What does it run on?
- What happens next?
XENON Overview

Industries we work in:

- Scientific / Academic Research
- Oil and Gas
- Cloud
- Defence
- Education
- Broadcast
- Finance
- Telecommunication
XENON Solutions

• Visual Workstations

XENON’s Nitro visual workstations: Performance and Reliability for the most demanding graphics, engineering, digital arts workloads and optimised for MATLAB.

• GPU Computing

High performance acceleration solutions for MATLAB leveraging NVIDIA Tesla technology and the CUDA ecosystem

• Virtualisation

End-to-end virtualisation solutions for compute, storage, networking, and desktop. Server and VDI solutions from high density servers to GPU enabled workstations and thin/zero client solutions.
XENON Milestones

Delivering world-class high performance computing solutions

1998
Telstra Digital Video Network
Supplied Video Server equipment

2000
Animal Logic
Render Farm delivered for the “Matrix” movie trilogy

2004
Tata Sky DTH India
Supplied Video Logging solution

2005
Thales Defence
Supplied Image Generators for the ASLAV simulators

2007
Victorian Partnership for Advance Computing
Supplied HPC Cluster for Advance Computing

2009
CSIRO
(10th in Green500 ranking)
Supplied Australia’s first GPU Cluster (“Bragg”)

2010
Australian Stock Exchange (ASX)
Supplied Infiniband equipment

2011
FOX Broadcasting USA
Supplied Video Logging solution

2012
Fujitsu
Supplied Infiniband Network for Australia’s largest Supercomputer (“Raijin”)

2013
Digital Cinema Systems
Upgraded 150 Independent Cinemas in ANZ region

2014
University of Queensland
FlashLite HPC cluster system for high data throughput

2015
Thales Defence
Delivered High Fidelity Solution for Australian Army Tiger Helicopter Simulator

2015
Futuris
HPC Linux Cluster in a Windows environment for crash test simulations
HPC GPU Cluster Bragg

Large scale GPU deployment

Designed, delivered, and installed by XENON Systems

• 384x Tesla K20 GPUs
• 384 GPUs = 958,464 Thread Processors
• 2048x 2GHz Intel Xeon E5-2650 Cores
• 16.4TB DDR3-1600 System Memory
• 128TB SATA2 Local System Storage
• InfiniBand Interconnect FDR10 40Gb/s
• Linpack Result: 335Tflops (Double Precision)
• Peak power usage: 115 kW
• Currently #297 in Top500 and #24 in Green500 (Nov. 2015)
What’s in an image?

An image says more than 1000 words…but what does it say?
What’s in an image?

*) Label determined based on inference by MATLAB using MatConvNet (see demo script below)
What’s in an image?
What’s in an image?
What’s in an image?

- Beagle dog
- Bell peppers
- Zebra
- Fish
What’s in an image?
How does MATLAB do it?  Looking under the hood

• Design a Deep Neural Network
• Train the network
• Present new images to the network
• Be prepared to be surprised…

Every network is only as good as its training.
What is required?

• System with NVIDIA GPU
• OS (Ubuntu 14.04 is a commonly used platform)
• NVIDIA drivers
• NVIDIA cuDNN library
• MatConvNet library: MATLAB toolbox implementing Convolutional Neural Networks (CNNs) for computer vision applications
• MATLAB
• MATLAB Parallel Computing Toolbox™, Computer Vision System Toolbox™ and Statistics and Machine Learning Toolbox™
• A little bit of MATLAB code…
Object Recognition in 7 lines of MATLAB Code

% Download pretrained network from MatConvNet repository
urlwrite('http://www.vlfeat.org/matconvnet/models/imagenet-vgg-f.mat', 'imagenet-vgg-f.mat');

% Load the network
cnnModel.net = load('imagenet-vgg-f.mat');

% Set up MatConvNet
run(fullfile('/opt/matconvnet-1.0-beta20','matlab','vl_setupnn.m'));

% choose a test image and display it
im='pet_images/bell-peppers.jpg';
imshow(im);

% Predict its content using ImageNet trained vgg-f CNN model
label = cnnPredict(cnnModel,img);
title(label,'FontSize',20)

NVIDIA Deep Learning SDK
High Performance GPU-Acceleration for Deep Learning

DEEP LEARNING FRAMEWORKS

- Caffe
- Chainer
- DL4J
- Mocha.jl
- KERAS
- CNTK
- mxnet
- Purine
- MatConvNet
- MINERVA
- OpenDeep
- Pylearn2
- TensorFlow
- Theano

DEEP LEARNING
- cuDNN

MATH LIBRARIES
- cuBLAS
- cuSPARSE
- cuFFT

MULTI-GPU
- NCCL

COMPONENTS
- Image Classification
- Voice Recognition
- Language Translation
- Recommendation Engines
- Sentiment Analysis

DEEP LEARNING
- Object Detection
- ImageNet
- Speech and Audio
- Image Classification
- Speech Recognition
- Natural Language Processing

MULTI-GPU
- NVIDIA Deep Learning SDK
- High Performance GPU-Acceleration for Deep Learning

MATH LIBRARIES
- cuBLAS
- cuSPARSE
- cuFFT

MULTI-GPU
- CUDA
- NCCL
NVIDIA cuDNN

Building blocks for accelerating deep neural networks on GPUs

- High performance deep neural network training
- Accelerates Deep Learning: Caffe, CNTK, Tensorflow, Theano, Torch
- Performance continues to improve over time

“NVIDIA has improved the speed of cuDNN with each release while extending the interface to more operations and devices at the same time.”

— Evan Shelhamer, Lead Caffe Developer, UC Berkeley
What’s new in cuDNN 5?

Pascal GPU, RNNs, Improved Performance

LSTM recurrent neural networks deliver up to 6x speedup in Torch

Improved performance:

- Deep Neural Networks with 3x3 convolutions, like VGG, GoogleNet and ResNets
- 3D Convolutions
- FP16 routines on Pascal GPUs

5.9x

Speedup for char-rnn RNN Layers

2.8x

Speedup for DeepSpeech 2 RNN Layers

Performance relative to torch-rnn
(https://github.com/jcjohnson/torch-rnn)
Char-rnn: https://github.com/karpathy/char-rnn
NVIDIA DIGITS
Interactive Deep Learning GPU Training System

Process Data
Configure DNN
Monitor Progress
Visualize Layers

developer.nvidia.com/digits
NVIDIA DIGITS
Improves Deep Learning Training Productivity

- Train neural network models with Torch support (preview)
- Save time by quickly iterating to identify the best model
- Manage multiple jobs easily to optimize use of system resources
- Active open source project with valuable community contributions

New Results Browser!
Preview DIGITS Future

Object Detection Workflow

- Object Detection Workflows for Automotive and Defense
- Targeted at Autonomous Vehicles, Remote Sensing

developer.nvidia.com/digits
What does it run on?

- XENON workstations with NVIDIA GPUs
- XENON DEVCUBE (see it in action at our stand!)
- XENON Radon rack servers
- 1U high density servers (up to 4 GPUs)
- 4U 8-GPU servers
- Custom configurations for your requirements
- and…
NVIDIA DGX-1 - The World’s First Deep Learning Supercomputer in a Box

System Specifications

The NVIDIA DGX-1 system specifications include:
Up to 170 teraflops of half-precision (FP16) peak performance
Eight Tesla P100 GPU accelerators, 16GB memory per GPU
NVLink Hybrid Cube Mesh
7TB SSD DL Cache
Dual 10GbE, Quad InfiniBand 100Gb networking
3U - 3200W

SYSTEM SPECIFICATIONS

- **GPUs**: 8x Tesla GP100
- **TFLOPS**: 42.5 Tflops double precision (FP64), 85 Tflops single precision (FP32), 170 Tflops half precision (FP16)
- **GPU Memory**: 16 GB per GPU
- **CPU**: Dual 16-core Intel Xeon E5-2698 v3 2.3 GHz
- **NVIDIA CUDA Cores**: 28672
- **System Memory**: 512 GB 2133 MHz DDR4 LRDIMM
- **Storage**: 4x 1.92 TB SSD RAID 0
- **Network**: Dual 10 GbE, 4 IB EDR
- **Software**: Ubuntu Server Linux OS, DGX-1 Recommended

© 2016 XENON Systems Pty Ltd. All rights reserved.
Pascal: The next generation GPU architecture

<table>
<thead>
<tr>
<th>GPU Architecture</th>
<th>NVIDIA Fermi</th>
<th>NVIDIA Kepler</th>
<th>NVIDIA Maxwell</th>
<th>NVIDIA Pascal</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPU Process</td>
<td>40nm</td>
<td>28nm</td>
<td>28nm</td>
<td>16nm</td>
</tr>
<tr>
<td>Maximum Transistors</td>
<td>3.54 Billion (GTX 690)</td>
<td>7.08 Billion (Titan Z)</td>
<td>8.00 Billion (Titan X)</td>
<td>7.2 B (1080) 15.3 B (P100)</td>
</tr>
<tr>
<td>Maximum Die Size</td>
<td>294mm²</td>
<td>561mm²</td>
<td>601mm²</td>
<td>610mm²</td>
</tr>
<tr>
<td>Stream Processors Per Compute Unit</td>
<td>32 SPs</td>
<td>192 SPs</td>
<td>128 SPs</td>
<td>64 SPs</td>
</tr>
<tr>
<td>Maximum CUDA Cores</td>
<td>512 CCs</td>
<td>2880 CCs</td>
<td>3072 CCs</td>
<td>3584 CCs</td>
</tr>
<tr>
<td>FP32 Compute</td>
<td>2.08 TFLOPs(Tesla)</td>
<td>5.04 TFLOPs (Tesla)</td>
<td>6.84 TFLOPs (Tesla)</td>
<td>~10.6TFLOPs (Tesla)</td>
</tr>
<tr>
<td>FP64 Compute</td>
<td>0.66 TFLOPs (Tesla)</td>
<td>1.68 TFLOPs (Tesla)</td>
<td>0.21 TFLOPs (Tesla)</td>
<td>5.3 TFLOPs(Tesla)</td>
</tr>
<tr>
<td>Maximum VRAM</td>
<td>1.5 GB GDDR5</td>
<td>12 GB GDDR5</td>
<td>24 GB GDDR5</td>
<td>16 GB HBM2</td>
</tr>
<tr>
<td>Maximum Bandwidth</td>
<td>192 GB/s</td>
<td>336 GB/s</td>
<td>336 GB/s</td>
<td>720 GB/s</td>
</tr>
</tbody>
</table>
Thank You

XENON Systems Pty Ltd
www.xenon.com.au

P +61 3 9549 1115 | F +61 3 9549 1199 | E info@xenon.com.au

A member of the XENON Technology Group
www.xtg.com.au