GPU TECHNOLOGY CONFERENCE

April 4-7, 2016 | Silicon Valley

S6474 - FROM WORKSTATION TO EMBEDDED: ACCELERATED DEEP LEARNING ON NVIDIA JETSON™ TX1

Julie Bernauer, 4/5/2016

PRESENTED BY NVIDIA.

GPU COMPUTING

GPU COMPUTING

ACCELERATED COMPUTING: CUDA

Framework to Program NVIDIA GPUs

A simple sum of two vectors (arrays) in C

```
void vector_add(int n, const float *a, const float *b, float *c)
{
  for( int idx = 0 ; idx < n ; ++idx )
    c[idx] = a[idx] + b[idx];
}
```

GPU friendly version in CUDA

```
__global__ void vector_add(int n, const float *a, const float *b, float *c)
{
  int idx = blockIdx.x*blockDim.x + threadIdx.x;
  if( idx < n )
    c[idx] = a[idx] + b[idx];
}
```

GPU ACCELERATED LIBRARIES

"Drop-in" Acceleration for Your Applications

Linear Algebra
FFT, BLAS,
SPARSE, Matrix, cuSolver

NVIDIA cuFFT, cuBLAS, cuSPARSE

CULA tools

MAGMA

CUSP

Numerical & Math
RAND, Statistics

NVIDIA Math Lib

ArrayFire

NVIDIA cuRAND

Data Struct. & AI
Sort, Scan, Zero Sum

Thrust

GPU AI – Board Games

GPU AI – Path Finding

Visual Processing
Image & Video

NVIDIA NPP

NVIDIA Video Encode

Sundog Software

DEEP NEURAL NETWORKS AND GPUS

ACCELERATING INSIGHTS

"Now You Can Build Google's $1M Artificial Brain on the Cheap"

WIRED

GOOGLE DATACENTER

1,000 CPU Servers
2,000 CPUs • 16,000 cores

600 kWatts
$5,000,000

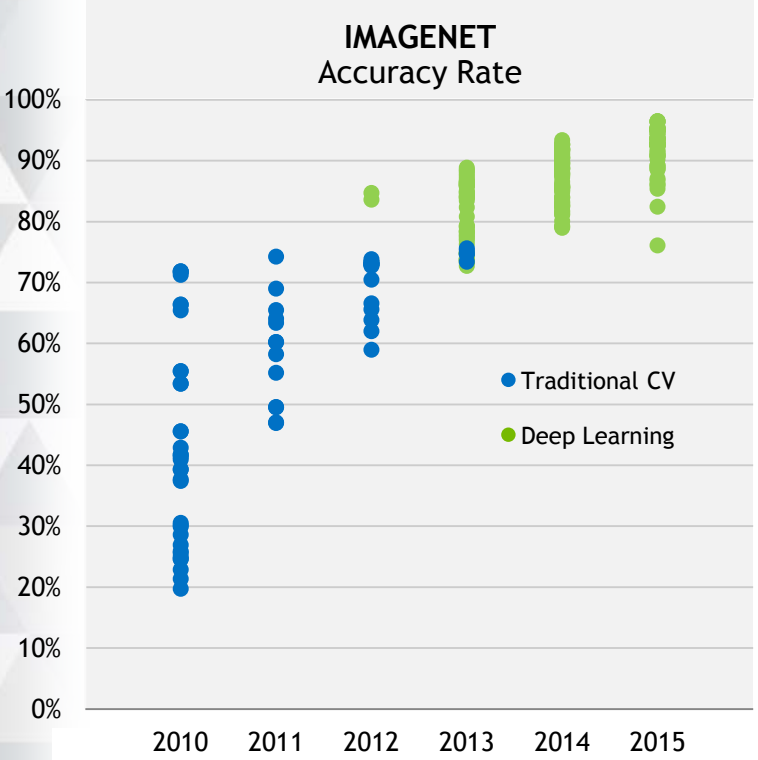
STANFORD AI LAB

3 GPU-Accelerated Servers
12 GPUs • 18,432 cores

4 kWatts
$33,000

Deep learning with COTS HPC systems, A. Coates, B. Huval, T. Wang, D. Wu, A. Ng, B. Catanzaro ICML 2013

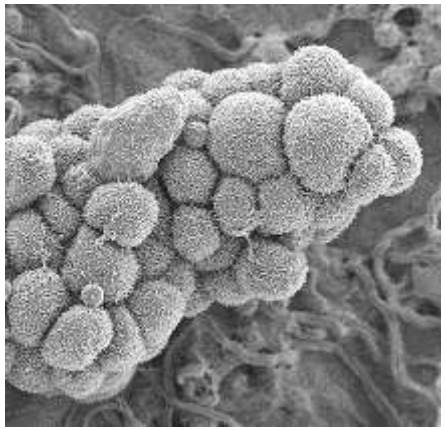
MODERN AI

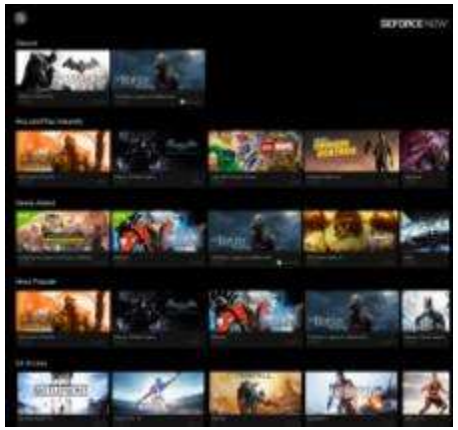


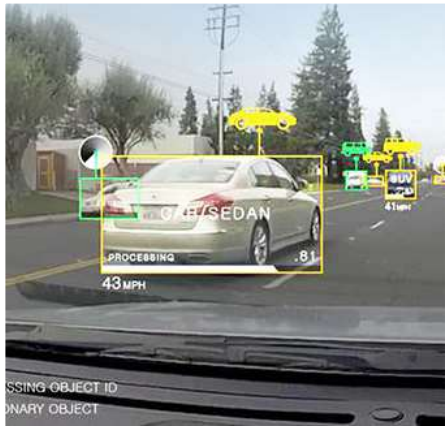
2012: GOOGLE BRAIN

2016: AlphaGO

DEEP LEARNING EVERYWHERE







INTERNET & CLOUD

Image Classification
Speech Recognition
Language Translation
Language Processing
Sentiment Analysis
Recommendation

MEDICINE & BIOLOGY

Cancer Cell Detection
Diabetic Grading
Drug Discovery

MEDIA & ENTERTAINMENT

Video Captioning
Video Search
Real Time Translation

SECURITY & DEFENSE

Face Detection
Video Surveillance
Satellite Imagery

AUTONOMOUS MACHINES

Pedestrian Detection
Lane Tracking
Recognize Traffic Sign

NVIDIA GPU: THE ENGINE OF DEEP LEARNING

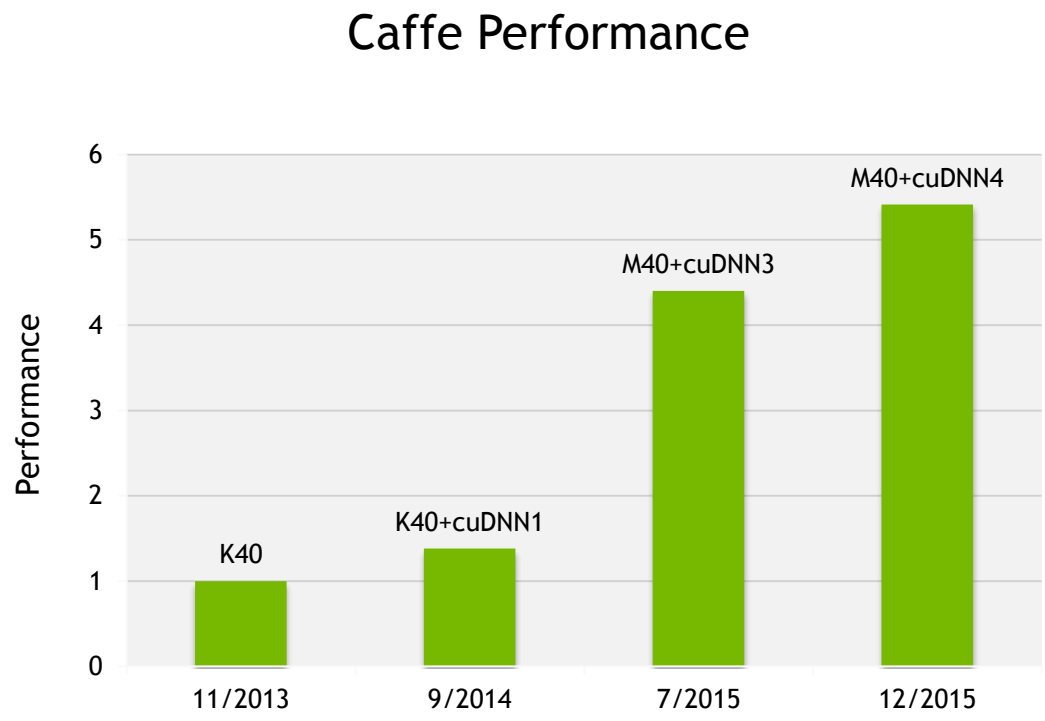
WATSON	CHAINER	THEANO	MATCONVNET
IBM	Preferred Networks	Université de Montréal	UNIVERSITY OF OXFORD

TENSORFLOW	CNTK	TORCH	CAFFE
Google	Microsoft	facebook	Berkeley UNIVERSITY OF CALIFORNIA

NVIDIA CUDA
ACCELERATED COMPUTING PLATFORM

ACCELERATING DEEP LEARNING: CUDNN

- GPU-accelerated Deep Learning subroutines
- High performance neural network training
- Accelerates Major Deep Learning frameworks: Caffe, Theano, Torch



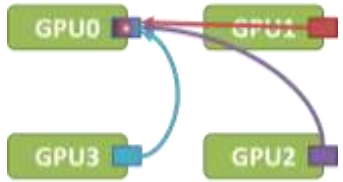
AlexNet training throughput based on 20 iterations,
CPU: 1x E5-2680v3 12 Core 2.5GHz. 128GB System Memory, Ubuntu 14.04

developer.nvidia.com/cudnn

MULTI-GPU COMMUNICATION: NCCL

Collective library

- Research library of accelerated collectives that is easily integrated and topology-aware so as to improve the scalability of multi-GPU applications



- Pattern the library after MPI's collectives
- Handle the intra-node communication in an optimal way
- Provide the necessary functionality for MPI to build on top to handle inter-node

github.com/NVIDIA/nccl

NCCL EXAMPLE

All-reduce

```
#include <nccl.h>
ncclComm_t comm[4];
ncclCommInitAll(comm, 4, {0, 1, 2, 3});
foreach g in (GPUs) { // or foreach thread
  cudaSetDevice(g);
  double *d_send, *d_recv;
  // allocate d_send, d_recv; fill d_send with data
  ncclAllReduce(d_send, d_recv, N, ncclDouble, ncclSum, comm[g], stream[g]);
  // consume d_recv
}
```

NVIDIA DEEP LEARNING SDK

High Performance GPU-Acceleration for Deep Learning

Image Classification

Object Detection

COMPUTER VISION

Voice Recognition

Language Translation

SPEECH AND AUDIO

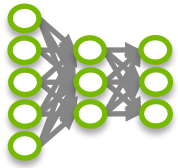
Recommendation Engines

Sentiment Analysis

NATURAL LANGUAGE PROCESSING

Caffe Chainer DL4J Deeplearning4j Mocha.jl julia KERAS Microsoft CNTK MatConvNet mxnet MINERVA Purine OpenDeep Pylearn2 TensorFlow theano torch

DEEP LEARNING FRAMEWORKS



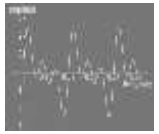
cuDNN

DEEP LEARNING

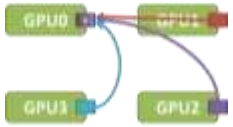
cuBLAS

cuSPARSE

cuFFT



MATH LIBRARIES

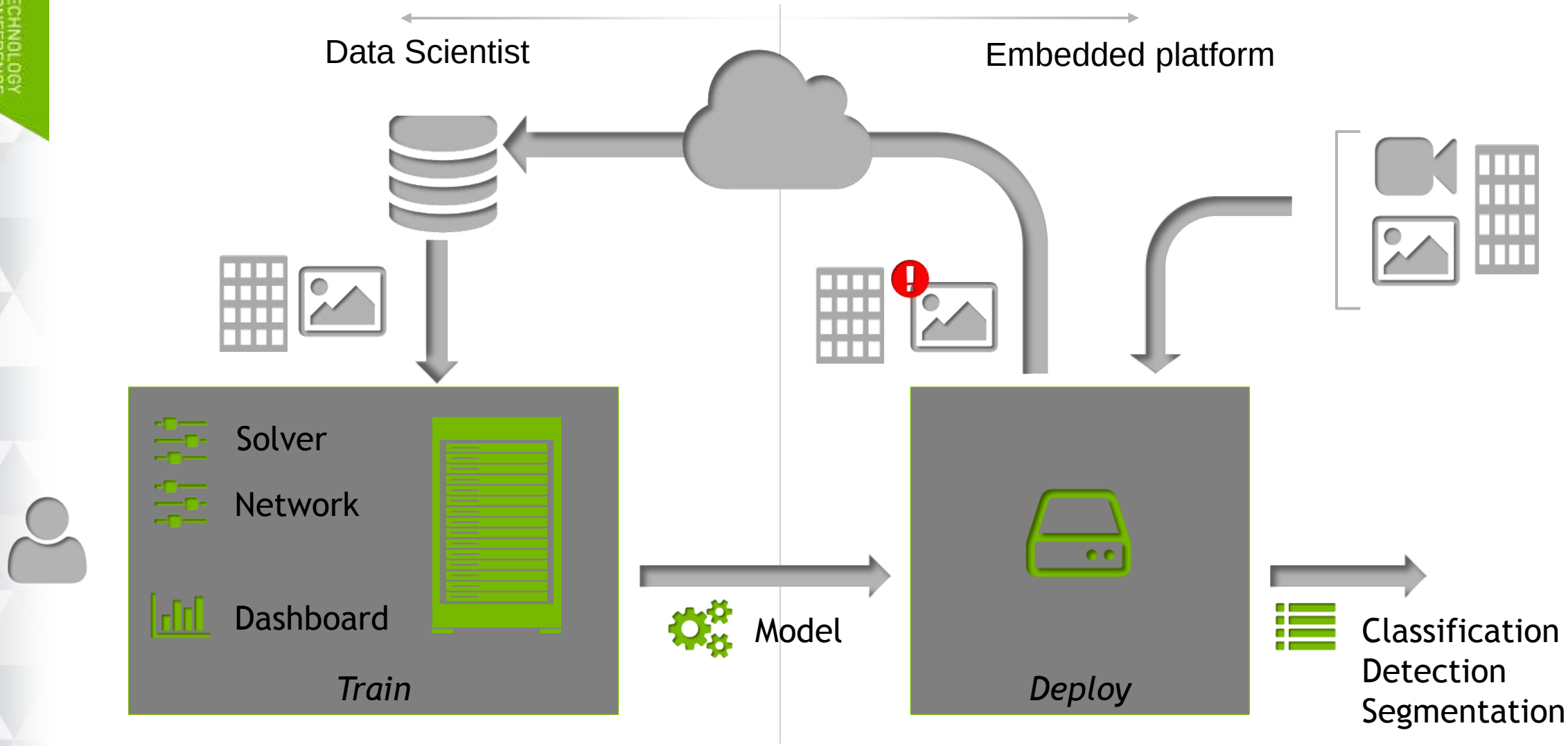


NCCL

MULTI-GPU

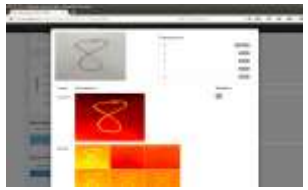
PLATFORM

AN END-TO-END SOLUTION

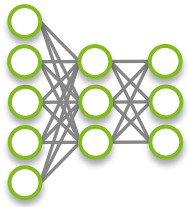


CUDA FOR DEEP LEARNING DEVELOPMENT

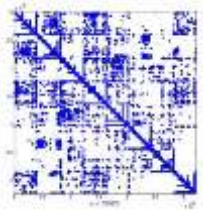
DEEP LEARNING SDK



DIGITS

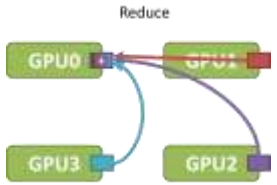


cuDNN



cuSPARSE

cuBLAS



NCCL

TITAN X

DEVBOX

GPU CLOUD

阿里云计算 Alibaba Cloud Computing

amazon web services

IBM SOFTLAYER

Microsoft Azure

TESLA FOR HYPERSCALE DATACENTERS

HYPERSCALE SUITE

Deep Learning SDK	GPU REST Engine	GPU Accelerated FFmpeg	Image Compute Engine	GPU support in Mesos

TESLA M40

POWERFUL: Fastest Deep Learning Performance

TESLA M4

LOW POWER: Highest Hyperscale Throughput

JETSON FOR INTELLIGENT MACHINES

JETSON SDK

IMAGE ARITHMETIC

- Absolute Difference
- Accumulate Image
- Accumulate Squared
- Accumulate Weighted
- Add / Subtract / Multiply
- Channel Combine
- Channel Extract
- Color Convert
- CopyImage
- Convert Depth
- Magnitude
- Not / Or / And / Xor
- Phase
- Table Lookup
- Threshold

FLOW & DEPTH

- Motion Flow
- Optical Flow (LK)
- Semi-Global Matching
- Stereo Block Matching

GEOMETRIC TRANSFORMS

- Affine Warp
- Warp Perspective
- Flip Image
- Remap
- Scale Image

FILTERS

- BoxFilter
- Convolution
- Dilation Filter
- Erosion Filter
- Gaussian Filter
- Gaussian Pyramid
- Laplacian3x3
- Median Filter
- Scharr3x3
- Sobel 3x3

FEATURES

- Canny Edge Detector
- Fast Corners
- Fast Track
- Harris Corners
- Harris Track
- Hough Circles
- Hough Lines

ANALYSIS

- Histogram
- Histogram Equalization
- Integral Image
- Mean Std Deviation
- Min Max Locations

JETSON TX1

GPU	1 TFLOP/s 256-core Maxwell
CPU	64-bit ARM A57 CPUs
Memory	4 GB LPDDR4 | 25.6 GB/s
Storage	16 GB eMMC
Size	50mm x 87mm
Power	Under 10W

NVIDIA DEEP LEARNING PLATFORM

DL FRAMEWORK (CAFFE, CNTK,TENSORFLOW, THEANO, TORCH…)

DEEP LEARNING SDK

TITAN X - DEVELOPERS

TESLA - DEPLOYMENT

amazon web services

Microsoft Azure

IBM SOFTLAYER

阿里云计算 Alibaba Cloud Computing

AUTOMOTIVE - DRIVEPX

EMBEDDED - JETSON

USING THE GPU FOR DEEP LEARNING

DIGITS™

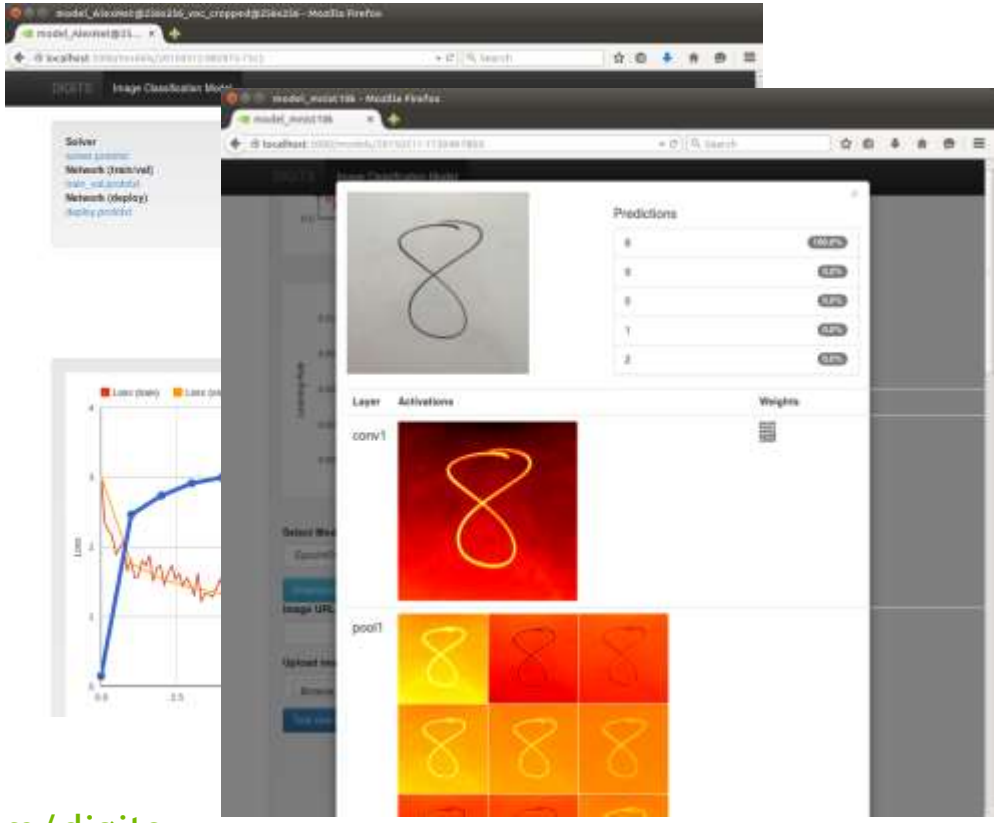
Interactive Deep Learning GPU Training System

Quickly design the best deep neural network (DNN) for your data

Train on multi-GPU (automatic)

Visually monitor DNN training quality in real-time

Manage training of many DNNs in parallel on multi-GPU systems



developer.nvidia.com/digits

FOR THE DEVELOPER: DIGITS™ DEVBOX

Fastest Deskside Deep Learning Training Box

Four TITAN X GPUs with 12GB of memory per GPU

1600W Power Supply Unit

Ubuntu 14.04

NVIDIA-qualified driver

NVIDIA® CUDA® Toolkit 7.0

NVIDIA® DIGITS™ SW

Caffe, Theano, Torch, BIDMach

FOR THE DATACENTER: MULTI-GPU SERVERS

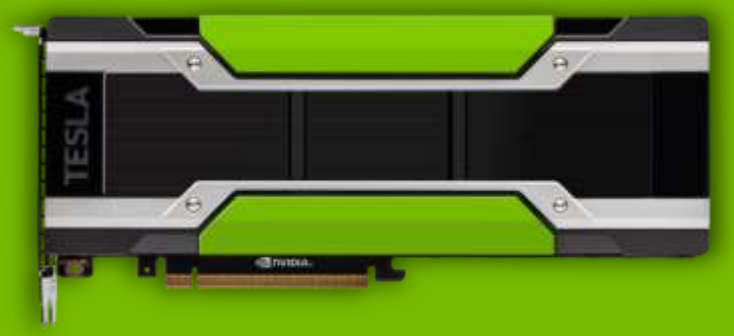
For accelerated training

Facebook

Big Sur Opencompute server

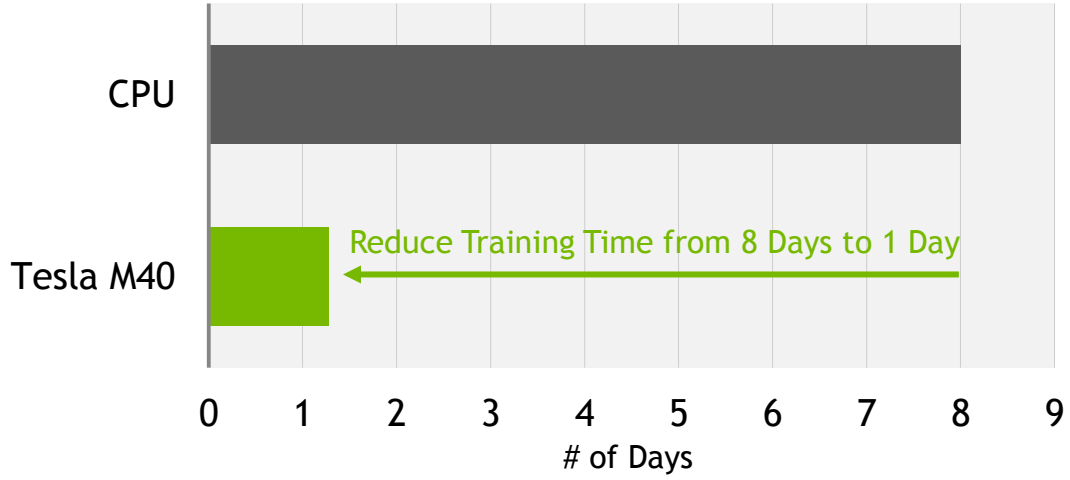
TESLA M40

World's Fastest Accelerator for Deep Learning



8x Faster

Caffe Performance

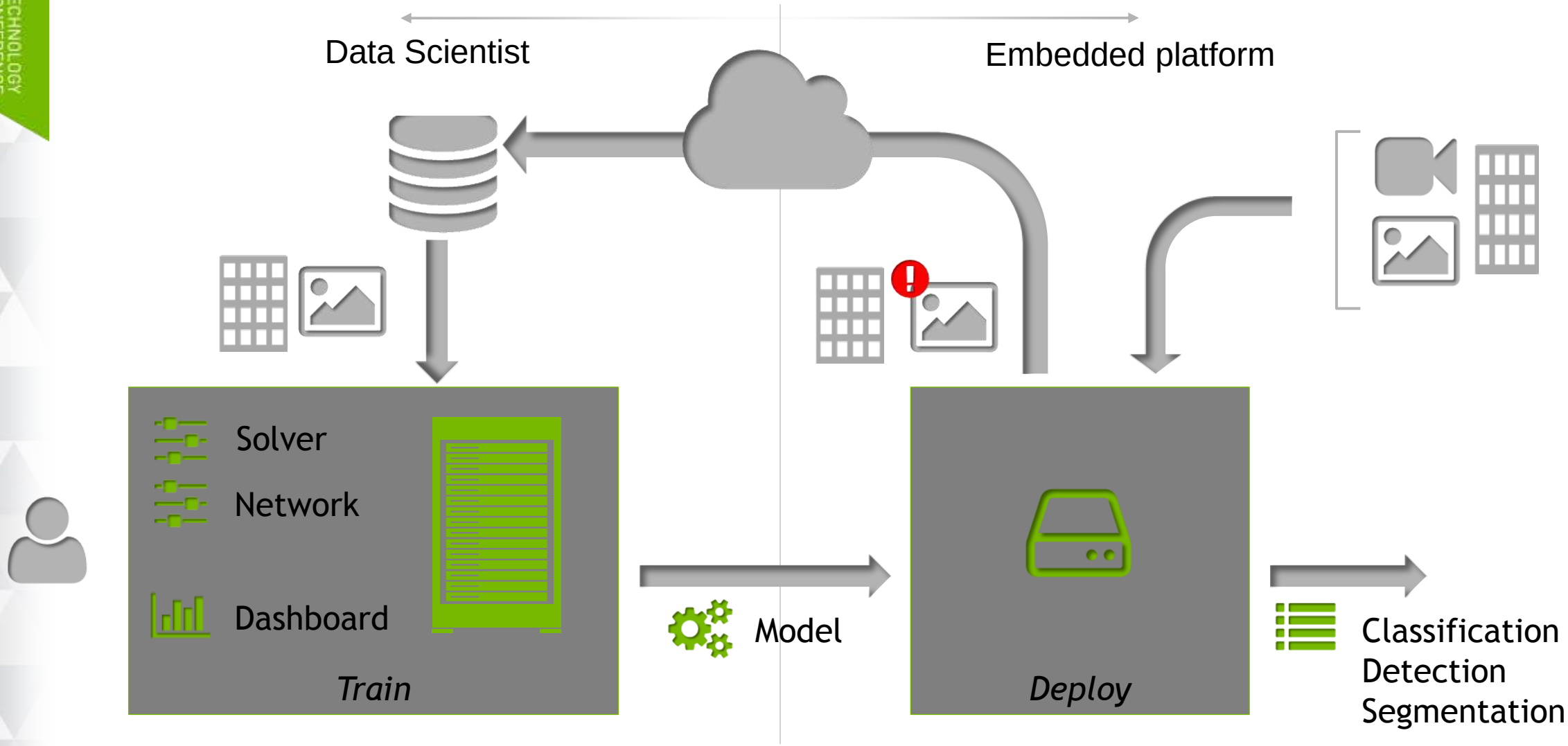


CUDA Cores	3072
Peak SP	7 TFLOPS
GDDR5 Memory	12 GB
Bandwidth	288 GB/s
Power	250W

Caffe Benchmark: AlexNet training throughput based on 20 iterations, CPU: E5-2697v2 @ 2.70GHz. 64GB System Memory, CentOS 6.2

THE DATA SCIENTIST JOURNEY MADE EASIER WITH GPUS

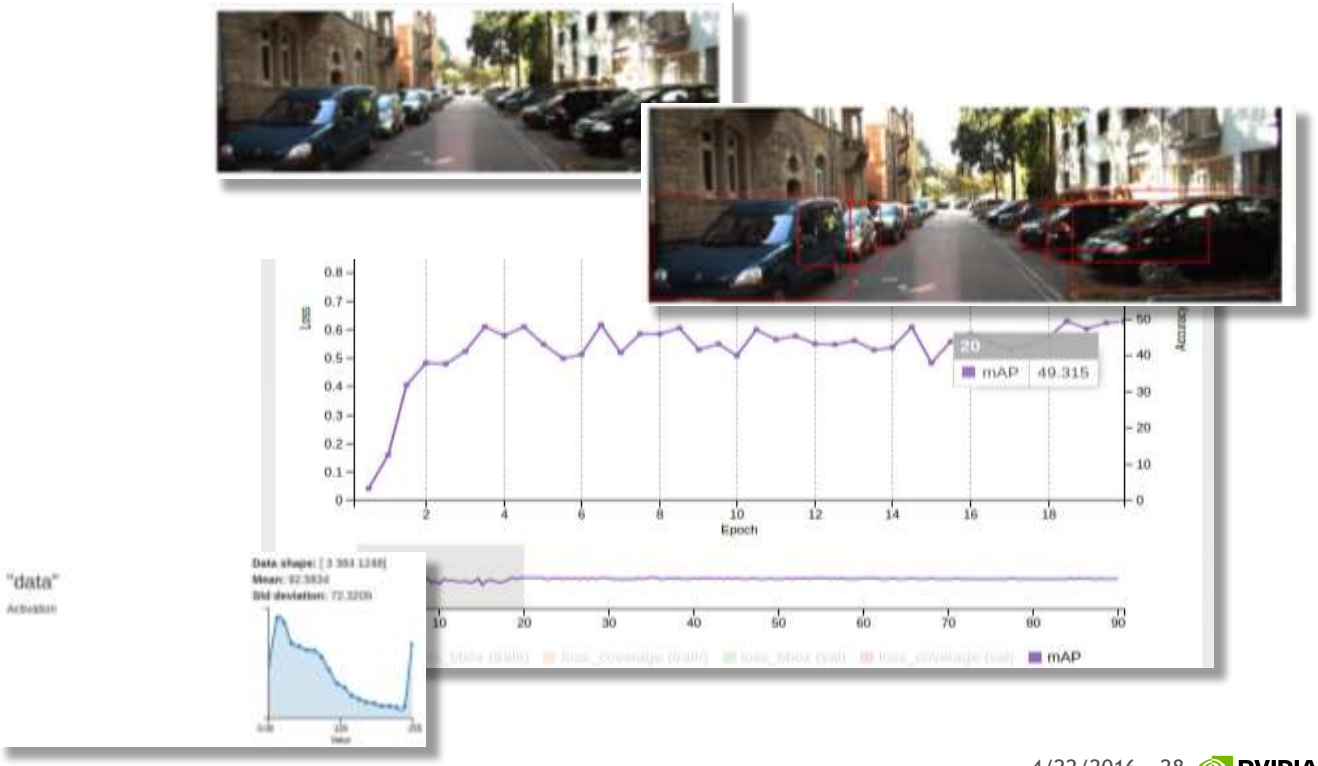
AN END-TO-END SOLUTION



SETTING UP DEEP LEARNING APPLICATIONS

Steps

- Data labelling / Curation
- Hyperparameter search
- Training
- Deployment



TRAINING EXAMPLE
DIGITS ON THE AMAZON CLOUD

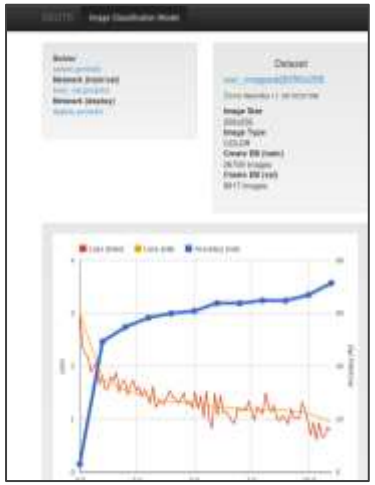

agray@sabox:~$

CLASSIFICATION EXAMPLE

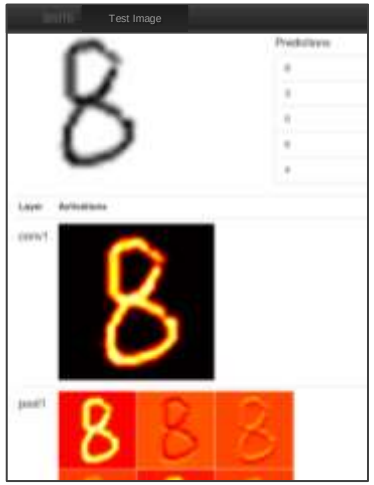
Process Data

Configure DNN

Monitor Progress



Visualize Layers



Come to the lab:

Tuesday 3:00pm

L6139

A Tutorial on More Ways to Use DIGITS

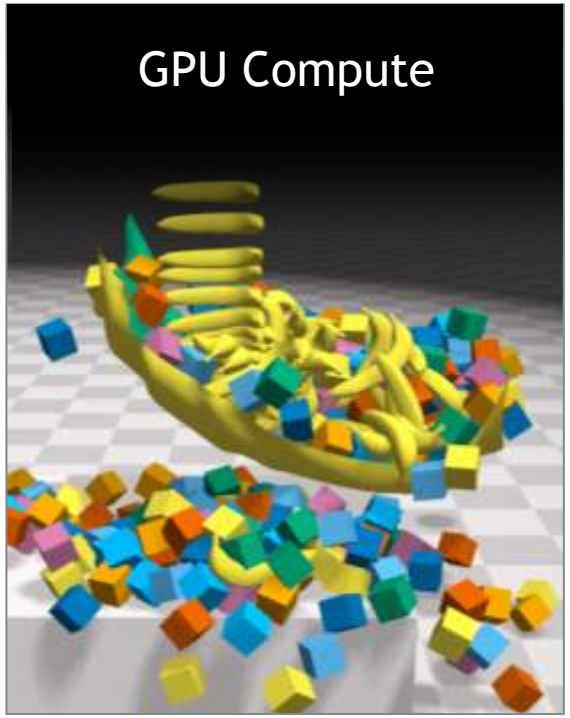
EMBEDDED DEPLOYMENT
JETSON TX1 DEVKIT

JETSON LINUX SDK

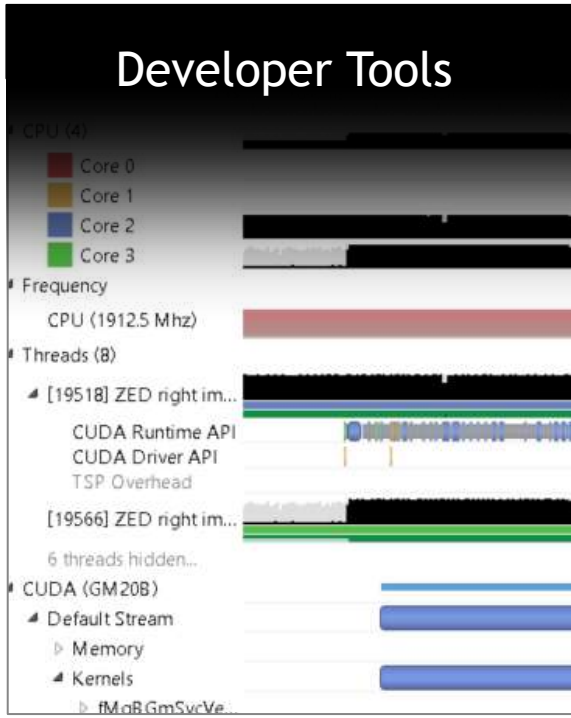
Graphics

Deep Learning and Computer Vision

GPU Compute



Developer Tools



OpenGL ES. OpenGL. Vulkan

cuDNN NVIDIA VISIONWORKS

NVIDIA CUDA

NVIDIA Nsight Eclipse Edition NVTX NVIDIA Tools eXtension

Debugger | Profiler | System Trace

EMBEDDED DEPLOYMENT

On Jetson TX1

Inference at 258 img/s

No need to change code

Simply compile Caffe and copy a trained `.caffemodel` to TX1

GOOGLENET PERFORMANCE

FP32 and FP16 inference

Network: GoogLeNet	Batch Size	Tegra X1 (FP32)	Tegra X1 (FP16)
Inference Performance	1	33 img/sec	33 img/sec
Power		5.0 W	4.0 W
Performance/Watt		6.5 img/sec/W	8.3 img/sec/W
Inference Performance	128 (Titan X) 64 (Tegra X1)	52 img/sec	75 img/sec
Power		5.9 W	5.8 W
Performance/Watt		8.8 img/sec/W	12.8 img/sec/W

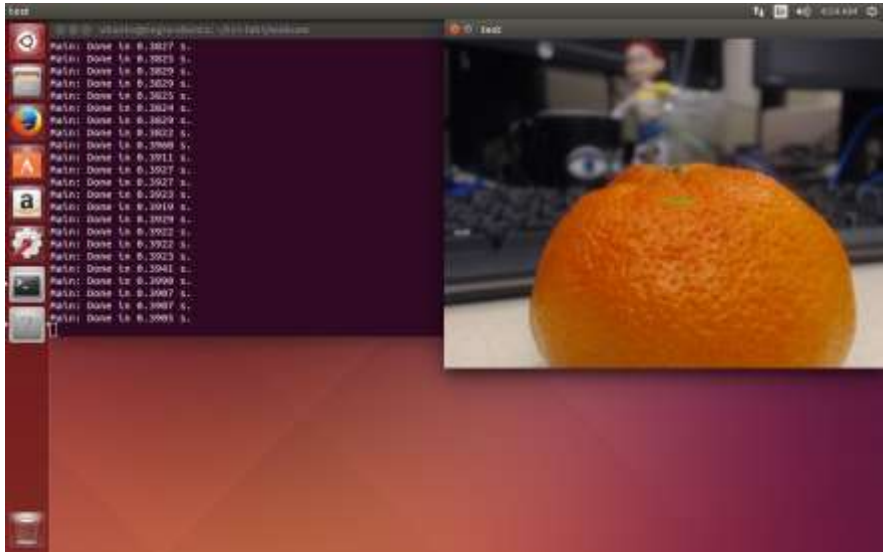
Server version (Tesla M40)
batch 1: ~130 img/sec - 120W / batch 128: ~850 img/sec – 225 W

CLASSIFICATION EXAMPLE

File Edit View Search Terminal Help

```
ubuntu@tegra-ubuntu:~$
```

CLASSIFICATION EXAMPLE



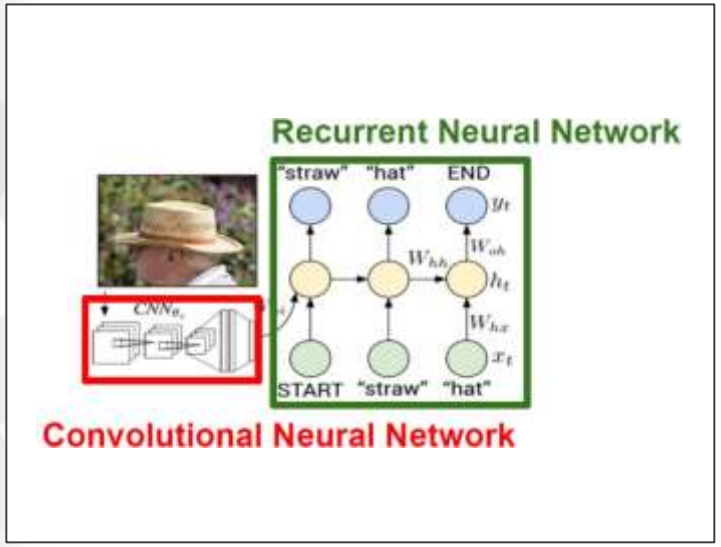
Come to the lab:

Tuesday 1:00pm

L6131 From Workstation to Embedded: Accelerated Deep Learning on NVIDIA Jetson TX1

OTHER USE CASES

IMAGE CAPTIONING



"Automated Image Captioning with ConvNets and Recurrent Nets"

—Andrej Karpathy, Fei-Fei Li

DRONES

RESOURCES

DEEP LEARNING AT GTC

Deep Learning at NVIDIA, Monday 4/4

11:00am: Accelerate Deep Learning with NVIDIA's Deep Learning Platform

12:00pm: Hangout -- The DIGITS Roadmap

1:00pm: From Workstation to Embedded: Accelerated Deep Learning on NVIDIA Jetson TX1

3:00pm: A Tutorial on More Ways to Use DIGITS

4:00pm: Hangout – cuDNN -- Features, Roadmap and Q&A

GPU TECHNOLOGY CONFERENCE

DEEP LEARNING AT GTC

Frameworks Hands-on Labs

Wednesday 4/6

1:00pm: Introduction to CNTK

2:00pm: Machine Learning Using TensorFlow

2:00pm: BIDMach Machine Learning Toolkit

3:30pm: Applied Deep Learning for Vision and Natural Language with Torch7

3:30pm: Caffe Hands-on Lab

Thursday 4/7

9:30am: Chainer Hands-on: Introduction To Train Deep Learning Model in Python

9:30am: Deep Learning With the Theano Python Library

1:00pm: IBM Watson Developers Lab

GPU TECHNOLOGY CONFERENCE

April 4-7, 2016 | Silicon Valley

THANK YOU

JOIN THE CONVERSATION

#GTC16

JOIN THE NVIDIA DEVELOPER PROGRAM AT developer.nvidia.com/join

PRESENTED BY

NVIDIA.