ACCELERATE DEEP LEARNING

WITH NVIDIA'S DEEP LEARNING PLATFORM | STEPHEN JONES | GTC’16
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
DEEP LEARNING REVOLUTIONIZING COMPUTING
Solves Problems Previously Unsolvable
A NEW COMPUTING MODEL

Traditional Computer Vision

Domain experts design feature detectors
Quality = patchwork of algorithms
Need CV experts and time

Deep Learning

DNN learn features from large data
Quality = data & training method
Needs lots of data and compute
DEEP LEARNING
The Next Innovation Network?

THE AI RACE IS ON

IMAGENET Accuracy Rate

IBM Watson Achieves Breakthrough in Natural Language Processing

Facebook Launches Big Sur

Baidu Deep Speech 2 Beats Humans

Google Launches TensorFlow

Toyota Invests $1B in AI Labs

Microsoft & U. Science & Tech, China Beat Humans on IQ
$500B DEEP LEARNING OPPORTUNITY

Deep Learning Total Revenue by Segment
World Markets: 2015-2024

Deep Learning Software Revenue by Industry
World Markets: 2015-2024

“The current cutting edge of deep learning processing platforms seems to be massively parallel GPU systems.”

— Tractica

# The Engine of Modern AI

## Education

<table>
<thead>
<tr>
<th>TORCH</th>
<th>CAFFE</th>
</tr>
</thead>
<tbody>
<tr>
<td>NYU</td>
<td>Facebook</td>
</tr>
<tr>
<td>Université de Montréal</td>
<td>University of Oxford</td>
</tr>
<tr>
<td>MIT</td>
<td>NUS</td>
</tr>
<tr>
<td>NYU</td>
<td>Microsoft</td>
</tr>
</tbody>
</table>

## Start-Ups

<table>
<thead>
<tr>
<th>CHAINER</th>
<th>DL4J</th>
<th>KERAS</th>
<th>OPENDEEP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preferred Networks</td>
<td>SKYMIND</td>
<td>SCHULTS LABORATORIES</td>
<td>VITRUVIAN</td>
</tr>
</tbody>
</table>

## NVIDIA Deep Learning Platform

*U. Washington, CMU, Stanford, TuSimple, NYU, Microsoft, U. Alberta, MIT, NYU Shanghai*
NVIDIA DEEP LEARNING PLATFORM

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

DEEP LEARNING SDK

TITAN X - DEVELOPERS

TESLA - DEPLOYMENT

AUTOMOTIVE - DRIVEPX

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

**VISION**
- Image Classification
- Object Detection

**SPEECH**
- Voice Recognition
- Language Translation

**BEHAVIOR**
- Recommendation Engines
- Sentiment Analysis

**DEEP LEARNING FRAMEWORKS**
- Caffe
- DL4J
- Mocha.jl
- Keras
- CNTK
- MXNet
- Theano
- Pylearn2
- TensorFlow
- Chainer

**DEEP LEARNING**
- cuDNN

**MATH LIBRARIES**
- cuBLAS
- cuSPARSE
- cuFFT

**MULTI-GPU**
- NCCL
NVIDIA Deep Learning SDK

Powerful developer tools and libraries for designing and deploying GPU-accelerated deep learning applications

- High Performance Deep Learning for NVIDIA GPUs
- Industry Vetted Deep Learning Algorithms
- Easily integrated into deep learning applications

developer.nvidia.com/deep-learning
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

developer.nvidia.com/cudnn
cuBLAS
Accelerated Linear Algebra for Deep Learning

GPU-accelerated Basic Linear Algebra Subroutines that delivers 6x to 17x faster performance than the latest MKL BLAS

- Accelerated Level 3 BLAS: SGEMM, SYMM, TRSM, SYRK
- Up to 7 TFlops Single Precision on a single M40
- Multi-GPU BLAS support available in cuBLAS-XT

developer.nvidia.com/cublas
NCCL

Accelerating Multi-GPU Communications

A topology-aware library of accelerated collectives to improve the scalability of multi-GPU applications

- Patterned after MPI’s collectives: includes all-reduce, all-gather, reduce-scatter, reduce, broadcast
- Optimized intra-node communication
- Supports multi-threaded and multi-process applications

github.com/NVIDIA/nccl
NVIDIA DIGITS
Making Deep Learning Accessible

An interactive development environment for training deep neural networks

- Prepare data quickly and easily for training
- Visualize network behavior
- Maximize training speed

developer.nvidia.com/digits
What’s new in DIGITS 3?

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!

developer.nvidia.com/digits
Preview DIGITS Future
Object Detection Workflow

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

Come see a live demo in the GTC Exhibit Hall!

developer.nvidia.com/digits
Deep Learning at GTC
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
Deep Learning at GTC
Frameworks Track, Wednesday 4/6

9:00am: Caffe: an Open Framework for Deep Learning
10:00am: TensorFlow: Scaling Up Machine Learning
2:00pm: Torch: A Flexible Platform for Deep Learning Research
3:00pm: Chainer: A Powerful, Flexible, and Intuitive Deep Learning Framework
4:00pm: Theano at a Glance: A Framework for Machine Learning
4:30pm: Deep Learning in Microsoft with CNTK

Monday, 4/4 at 3:00pm: MXNet: Flexible Deep Learning Framework from Distributed GPU Clusters to Embedded Systems
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

Monday 4/4
1:00pm: Train and Deploy Deep Learning for Vision, Natural Language and Speech Using MXNet
Deep Learning at GTC

Over 50 sessions on Deep Learning, highlights include --

Tuesday, 4/5

1:00pm: Distributed Deep Learning at Scale, Soumith Chintala, Research Engineer, Facebook AI Research

2:00pm: Generative Adversarial Networks, Ian Goodfellow, Senior Research Scientist, Google

3:00pm: Video Classification of Live Streams on Twitter's Periscope, Nicolas Koumchatzky, Engineer, Twitter

4:00pm: Training and Deploying Deep Neural Networks for Speech Recognition, Bryan Catanzaro, Senior Researcher, Baidu Research

Wednesday, 4/6, 9:30am: Deep Reinforcement Learning, Pieter Abbeel, Professor, UC Berkeley
Deep Learning at GTC

**Hangouts**

**Monday 4/4**

12:00pm: The DIGITS Roadmap

4:00pm: cuDNN—Features, Roadmap and Q&A

**Tuesday 4/5**

12:00pm: Dreaming Big: Scaling Up Deep Dream to Operate on Multi-Hundred Megapixel Images

**Wednesday 4/6**

10:00am: Deep Learning in Image and Video

1:00pm: Deep Learning Exploits Petabytes of DigitalGlobe GIS

1:00pm: Reinforcement Learning

**Thursday 4/7**

1:00pm: Large Vocabulary Speech Recognition with GPUs

1:00pm: NVIDIA Deep Learning Software
Deep Learning at GTC

**NLP/NLU**

S6515 - Listen, Attend and Spell, William Chan, PhD Candidate, Carnegie Mellon University

S6745 - VQA: Visual Question Answering, Aishwarya Agrawal, PhD Student, Virginia Tech

S6781 - Deep Neural Networks for Conversational Language Understanding, Kaheer Suleman, CTO, Maluuba Inc.

S6321 - How Deep Learning Works for Automated Customer Service, Chenghua (Kevin) Li, Chief Scientist of DNN Lab, JD.COM

H6127 - Hangout: Large Vocabulary Speech Recognition with GPUs, Yifang Xu, Senior Deep Learning Software Engineer, NVIDIA
Deep Learning at GTC

Speech

S6371 - Deep Convolutional Neural Networks for Spoken Dialect Classification of Spectrogram Images Using DIGITS, Nigel Cannings, Chief Technical Officer, Intelligent Voice Limited

S6383 - High Performance CTC Training for End-to-End Speech Recognition on GPU, Minmin Sun, GPU Architecture Engineer, NVIDIA

S6458 - A GPU-Based Cloud Speech Recognition Server for Dialog Applications, Alexei V. Ivanov, CTO, Verbumware Inc.

S6672 - Training and Deploying Deep Neural Networks for Speech Recognition, Bryan Catanzaro, Senior Researcher, Baidu Research (Highly-Rated Speaker)

S6515 - Listen, Attend and Spell, William Chan, PhD Candidate, Carnegie Mellon University

S6781 - Deep Neural Networks for Conversational Language Understanding, Kaheer Suleman, CTO, Maluuba Inc.
QUESTIONS?