DEEP LEARNING WITH GPUs

GEOINT 2015

Larry Brown Ph.D.

June 2015
Introducing NVIDIA
What is Deep Learning?
GPUs and Deep Learning
cuDNN and DiGiTS
Machine Learning & Data Analytics

...and a video!
INTRODUCING NVIDIA

I know what you’re thinking...

GAMING - SPECIAL FX
NVIDIA - INVENTOR OF THE GPU

NVIDIA Invented the GPU in 1999, with over 1 Billion shipped to date.

Initially a dedicated a graphics processor for PCs, the GPU’s computational power and energy efficiency led to it quickly being adopted for professional high performance computing and enterprise data analytics.

In 2007, NVIDIA launched the CUDA® programming platform, and opened up the general purpose parallel processing capabilities of the GPU.
NVIDIA PLATFORM

DESIGN and VISUALIZATION

MOBILE and EMBEDDED

HPC and DATA CENTER

GAMING

GPUs and SOCs

GRAPHICS CARDS

SYSTEMS

IP
USA - TWO FLAGSHIP SUPERCOMPUTERS

SUMMIT
150-300 PFLOPS Peak Performance
IBM POWER9 CPU + NVIDIA Volta GPU
NVLink High Speed Interconnect
>40 TFLOPS per Node, >3,400 Nodes
2017

SIERRA
> 100 PFLOPS Peak Performance
Powered by the NVIDIA GPU
BEYOND HPC TO BIG DATA ANALYTICS

<table>
<thead>
<tr>
<th>Boeing</th>
<th>Chevron</th>
<th>GE</th>
</tr>
</thead>
<tbody>
<tr>
<td>NREL</td>
<td>National Renewable Energy Laboratory</td>
<td>NASA</td>
</tr>
<tr>
<td>NCSA</td>
<td>Oak Ridge National Laboratory</td>
<td>Pacific Northwest National Laboratory</td>
</tr>
<tr>
<td>Siemens</td>
<td>Raytheon</td>
<td></td>
</tr>
<tr>
<td>Baidu</td>
<td>Coupons</td>
<td>Dropbox</td>
</tr>
<tr>
<td>Facebook</td>
<td>Google</td>
<td>Naver</td>
</tr>
<tr>
<td>Netflix</td>
<td>Palantir</td>
<td>Square</td>
</tr>
<tr>
<td>Walmart</td>
<td>Yahoo!</td>
<td>Yandex</td>
</tr>
</tbody>
</table>

Attendees at GPU Technology Conference, 2014
Researchers using Tesla GPUs are solving the world’s great scientific and technical challenges.

Using a supercomputer powered by 3,000 Tesla processors, University of Illinois scientists achieved a breakthrough in HIV research. Another research team from Baylor, Rice, MIT, Harvard and the Broad Institute used GPUs to map how the human genome folds within the nucleus of a cell.

These and other advances in science have been highlighted in top journals and are regularly showcased at GTC, our annual developer conference.
10X GROWTH IN GPU COMPUTING

- **2008**
  - 150,000 CUDA Downloads
  - 27 CUDA Apps
  - 60 Universities Teaching
  - 4,000 Academic Papers
  - 6,000 Tesla GPUs
  - 77 Supercomputing Teraflops

- **2015**
  - 3 Million CUDA Downloads
  - 319 CUDA Apps
  - 800 Universities Teaching
  - 60,000 Academic Papers
  - 450,000 Tesla GPUs
  - 54,000 Supercomputing Teraflops
GPU ACCELERATED COMPUTING
10X PERFORMANCE 5X ENERGY EFFICIENCY

CPU
Optimized for Serial Tasks

GPU Accelerator
Optimized for Parallel Tasks
What is Deep Learning?
“Machine Learning” is in some sense a rebranding of AI.

The focus is now on more specific, often perceptual tasks, and there are many successes.

Today, some of the world’s largest internet companies, as well as the foremost research institutions, are using GPUs for machine learning.
INDUSTRIAL USE CASES

...machine learning is pervasive
TRADITIONAL ML - HAND TUNED FEATURES

Images/video

Image
Vision features
Detection

Audio

Audio
Audio features
Speaker ID

Text

Text
Text features
Text classification, Machine translation, Information retrieval, ....

Slide courtesy of Andrew Ng, Stanford University
WHAT IS DEEP LEARNING?

Systems that learn to recognize objects that are important, without us telling the system explicitly what that object is ahead of time

Key components
- Task
- Features
- Model
- Learning Algorithm
THE PROMISE OF MACHINE LEARNING

ML Systems Extract Value From Big Data

facebook
350 millions images uploaded per day

Walmart
2.5 Petabytes of customer data hourly

YouTube
100 hours of video uploaded every minute
IMAGE CLASSIFICATION WITH DNNS

Training
- cars
- buses
- trucks
- motorcycles

Inference
- truck
IMAGE CLASSIFICATION WITH DNNS

Training

- cars
- buses
- trucks
- motorcycles

Typical training run

- Pick a DNN design
- Input 100 million training images spanning 1,000 categories
- *One week* of computation

Test accuracy

- If bad: modify DNN, fix training set or update training parameters
WHAT MAKES DEEP LEARNING DEEP?

Today’s Largest Networks

- 10 layers
- 1B parameters
- 10M images
- ~30 Exaflops
- ~30 GPU days

Human brain has trillions of parameters - only 1,000 more.
Deep Learning

- Don’t have to figure out the features ahead of time.
- Use same neural net approach for many different problems.
- Fault tolerant.
- Scales well.
CONVOLUTIONAL NEURAL NETWORKS

- Biologically inspired.

- Neuron only connected to a small region of neurons in layer below it called the *receptive field*.

- A given layer can have many convolutional filters/kernels. Each filter has the same weights across the whole layer.

- Bottom layers are convolutional, top layers are fully connected.

- Generally trained via supervised learning.
  
<table>
<thead>
<tr>
<th>Supervised</th>
<th>Unsupervised</th>
<th>Reinforcement</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>...ideal system automatically switches modes...</td>
</tr>
</tbody>
</table>
CONVOLUTIONAL NETWORKS BREAKTHROUGH


CNNS DOMINATE IN PERCEPTUAL TASKS

- Handwriting recognition MNIST (many), Arabic HWX (IDSIA)
- OCR in the Wild [2011]: StreetView House Numbers (NYU and others)
- Traffic sign recognition [2011] GTSRB competition (IDSIA, NYU)
- Asian handwriting recognition [2013] ICDAR competition (IDSIA)
- Pedestrian Detection [2013]: INRIA datasets and others (NYU)
- Volumetric brain image segmentation [2009] connectomics (IDSIA, MIT)
- Object Recognition [2012] ImageNet competition (Toronto)
- Scene Parsing [2012] Stanford bgd, SiftFlow, Barcelona datasets (NYU)
- Scene parsing from depth images [2013] NYU RGB-D dataset (NYU)
- Speech Recognition [2012] Acoustic modeling (IBM and Google)
- Breast cancer cell mitosis detection [2011] MITOS (IDSIA)
RECURRENT NEURAL NETWORK - RNN

AKA: “LSTM”

- Remembers prior state.
- Good for sequences.
- Predict next character given input text.
- Translate sentence between languages.
- Generate a caption for an image.
WHY IS DEEP LEARNING HOT **NOW?**

Three Driving Factors...

<table>
<thead>
<tr>
<th>Big Data Availability</th>
<th>New ML Techniques</th>
<th>Compute Density</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>facebook</strong></td>
<td>350 millions images uploaded per day</td>
<td>Deep Neural Networks</td>
</tr>
<tr>
<td><strong>Walmart</strong></td>
<td>2.5 Petabytes of customer data hourly</td>
<td></td>
</tr>
<tr>
<td><strong>YouTube</strong></td>
<td>100 hours of video uploaded every minute</td>
<td></td>
</tr>
</tbody>
</table>

ML systems extract value from Big Data
GPUs and Deep Learning
GPUs — THE HOT PLATFORM FOR MACHINE LEARNING

Image Recognition Challenge
1.2M training images • 1000 object categories

Hosted by

GPU Entries

Classification Error Rates

2010 2011 2012 2013 2014

0 20 40 60 80 100 120

28% 26% 16% 12% 7%

2010 2011 2012 2013 2014

0% 5% 10% 15% 20% 25% 30%
Deep learning with COTS HPC systems
A. Coates, B. Huval, T. Wang, D. Wu, A. Ng, B. Catanzaro
ICML 2013

“Now You Can Build Google’s $1M Artificial Brain on the Cheap”

GPUS MAKE DEEP LEARNING ACCESSIBLE

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 Image: Scaling up Image Recognition”
– Baidu: 5.98%, Jan. 13, 2015

“Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification”
– Microsoft: 4.94%, Feb. 6, 2015

– Google: 4.82%, Feb. 11, 2015
The theme of deep learning carried through our guest keynotes. Jeff Dean, senior fellow at Google, described how the company is using GPU-powered deep neural networks to bring greater levels of intelligence to image, text, and speech recognition. He also highlighted work done by the recently acquired Deep Mind. Using Atari video games, the researchers trained a network to not just classify, but take actions in an environment. Ultimately, the network beat a series of games and the work earned the cover of Nature magazine.

“We love GPU cards. We just use a lot of them.”

— Jeff Dean, Google
DATA AUGMENTATION

Augmentation expands your dataset

- Mirror images
- Distorted / blurred
- Rotations
- Color changes

Example from Baidu...
WHY ARE GPUs GOOD FOR DEEP LEARNING?

**Neural Networks** | **GPUs**
--- | ---
Inherently Parallel | ✓ | ✓
Matrix Operations | ✓ | ✓
FLOPS | ✓ | ✓

GPUs deliver --
- same or better prediction accuracy
- faster results
- smaller footprint
- lower power

![ImageNet Challenge Accuracy Graph](image)
### GPU ACCELERATION

**Training A Deep, Convolutional Neural Network**

<table>
<thead>
<tr>
<th>Batch Size</th>
<th>Training Time CPU</th>
<th>Training Time GPU</th>
<th>GPU Speed Up</th>
</tr>
</thead>
<tbody>
<tr>
<td>64 images</td>
<td>64 s</td>
<td>7.5 s</td>
<td>8.5X</td>
</tr>
<tr>
<td>128 images</td>
<td>124 s</td>
<td>14.5 s</td>
<td>8.5X</td>
</tr>
<tr>
<td>256 images</td>
<td>257 s</td>
<td>28.5 s</td>
<td>9.0X</td>
</tr>
</tbody>
</table>

- ILSVRC12 winning model: “Supervision”
- 7 layers
- 5 convolutional layers + 2 fully-connected
- ReLU, pooling, drop-out, response normalization
- Implemented with Caffe
- Dual 10-core Ivy Bridge CPUs
- 1 Tesla K40 GPU
- CPU times utilized Intel MKL BLAS library
- GPU acceleration from CUDA matrix libraries (cuBLAS)
DEEP LEARNING EXAMPLES

Image Classification, Object Detection, Localization, Action Recognition, Scene Understanding

Speech Recognition, Speech Translation, Natural Language Processing

Pedestrian Detection, Traffic Sign Recognition

Breast Cancer Cell Mitosis Detection, Volumetric Brain Image Segmentation
## GPU-Accelerated Deep Learning Frameworks

<table>
<thead>
<tr>
<th></th>
<th>CAFFE</th>
<th>TORCH</th>
<th>THEANO</th>
<th>CUDA-CONVNET2</th>
<th>KALDI</th>
</tr>
</thead>
<tbody>
<tr>
<td>cuDNN</td>
<td>R2</td>
<td>R2</td>
<td>R2</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>Multi-GPU</td>
<td>In Progress</td>
<td>Partial</td>
<td>Partial</td>
<td>✓</td>
<td>✓ (nnet2)</td>
</tr>
<tr>
<td>Multi-CPU</td>
<td>média</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓ (nnet2)</td>
</tr>
<tr>
<td>License</td>
<td>BSD-2</td>
<td>GPL</td>
<td>BSD</td>
<td>Apache 2.0</td>
<td>Apache 2.0</td>
</tr>
<tr>
<td>Interface(s)</td>
<td>Text-based definition files, Python, MATLAB</td>
<td>Python, Lua, MATLAB</td>
<td>Python</td>
<td>C++</td>
<td>C++, Shell scripts</td>
</tr>
<tr>
<td>Embedded (TK1)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

cuDNN and DiGiTS
HOW GPU ACCELERATION WORKS

Application Code

Compute-Intensive Functions
5% of Code
~ 80% of run-time

Rest of Sequential CPU Code

GPU

CPU
WHAT IS cuDNN?
cuDNN is a library of primitives for deep learning

Applications

Programming Languages
- Maximum Flexibility

Libraries
- "Drop-in" Acceleration

OpenACC Directives
- Easily Accelerate Applications
cuDNN is a library of primitives for deep learning

**ANALOGY TO HPC**

Application
- Fluid Dynamics
- Computational Physics

BLAS standard interface

Various CPU BLAS implementations
- Intel CPUs
- IBM Power

cuBLAS/NVBLAS
- Tesla
- Titan
- TK1
- TX1
DEEP LEARNING WITH cuDNN

cuDNN is a library of primitives for deep learning

Applications

Frameworks

Caffe  torch  Kaldi  theano

cuDNN

Tesla  TX-1  GPUs  Titan
cuDNN ROUTINES

- Convolutions - 80-90% of the execution time
- Pooling - Spatial smoothing
- Activation - Pointwise non-linear function
CONVOLUTIONS - THE MAIN WORKLOAD

- Very compute intensive, but with a large parameter space

<table>
<thead>
<tr>
<th>Number</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Minibatch Size</td>
</tr>
<tr>
<td>2</td>
<td>Input feature maps</td>
</tr>
<tr>
<td>3</td>
<td>Image Height</td>
</tr>
<tr>
<td>4</td>
<td>Image Width</td>
</tr>
<tr>
<td>5</td>
<td>Output feature maps</td>
</tr>
<tr>
<td>6</td>
<td>Kernel Height</td>
</tr>
<tr>
<td>7</td>
<td>Kernel Width</td>
</tr>
<tr>
<td>8</td>
<td>Top zero padding</td>
</tr>
<tr>
<td>9</td>
<td>Side zero padding</td>
</tr>
<tr>
<td>10</td>
<td>Vertical stride</td>
</tr>
<tr>
<td>11</td>
<td>Horizontal stride</td>
</tr>
</tbody>
</table>

- Layout and configuration variations

- Other cuDNN routines have straightforward implementations
cuDNN V2 - PERFORMANCE

CPU is 16 core Haswell E5-2698 at 2.3 GHz, with 3.6 GHz Turbo
GPU is NVIDIA Titan X
cuDNN EASY TO ENABLE

- Install cuDNN on your system
- Download CAFFE
- In CAFFE Makefile.config
  - uncomment USE_CUDNN := 1
- Install CAFFE as usual
- Use CAFFE as usual.

- Install cuDNN on your system
- Install Torch as usual
- Install cudnn.torch module
- Use cudnn module in Torch instead of regular nn module.
- cudnn module is API compatible with standard nn module. Replace nn with cudnn

CUDA 6.5 or newer required
DIGITS
Interactive Deep Learning GPU Training System

Data Scientists & Researchers:
- Quickly design the best deep neural network (DNN) for your data
- Visually monitor DNN training quality in real-time
- Manage training of many DNNs in parallel on multi-GPU systems

developer.nvidia.com/digits
Main Console

Create your dataset

Choose your database

Configure your Network

Choose your database

Choose a default network, modify one, or create your own

Start Training

Start training

DIGITS Workflow

Create your database

Configure your model

Start training
Machine Learning and Data Analytics
TRADITIONAL MACHINE LEARNING

For your many non-DL applications...

- Interactive environment for easily building and deploying ML systems.
- Holds records for performance on many common ML tasks, on single nodes or clusters.
- Uses Scala. Feels like SciPy or Matlab.
GPU ACCELERATION FOR GRAPH ANALYTICS

- Comms & Social networks
- Cyber pattern recognition
- Shortest path finding

PageRank: 19x Speedup

0.967

1 GPU vs 60 Nodes
- 280x vs optimized Spark
- 1440x vs Spark
- 3420x vs Hadoop
Thank you!