GPU ACCELERATED DEEP LEARNING WITH cuDNN

Larry Brown Ph.D.

March 2015
AGENDA

1. Introducing cuDNN and GPUs
2. Deep Learning Context
3. cuDNN V2
4. Using cuDNN
Introducing cuDNN and GPUs
HOW GPU ACCELERATION WORKS

Application Code

Compute-Intensive Functions

5% of Code
~ 80% of run-time

Rest of Sequential CPU Code
WHAT IS cuDNN?

cuDNN is a library of primitives for deep learning

Applications

Programming Languages  Libraries  OpenACC Directives

Maximum Flexibility  “Drop-in” Acceleration  Easily Accelerate Applications
cuDNN is a library of primitives for deep learning
DEEP LEARNING WITH cuDNN

cuDNN is a library of primitives for deep learning

Frameworks
- Caffe
- torch
- Kaldi
- theano

Applications

cuDNN

GPUs
- Tesla
- TX-1
- Titan
cuDNN V2 is focused on ...

Performance and,

Features

... for the deep learning practitioner!

Optimized for current and future GPUs
Deep Learning Context
“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.
MACHINE LEARNING USE CASES

...machine learning is pervasive

- Image Classification, Object Detection, Localization
- Face Recognition
- Speech & Natural Language Processing
- Medical Imaging & Interpretation
- Seismic Imaging & Interpretation
- Recommendation
WHY IS DEEP LEARNING HOT NOW?

THREE DRIVING FACTORS...

1 - Big Data Availability

- Facebook: 350 millions images uploaded per day
- Walmart: 2.5 Petabytes of customer data hourly
- YouTube: 100 hours of video uploaded every minute

2 - New ML Techniques

- Deep Neural Networks

3 - Compute Density

- GPUs

ML systems extract value from Big Data
DIFFERENT MODALITIES...SAME APPROACH

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
DEEP LEARNING ADVANTAGES

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.
WHAT IS DEEP LEARNING?

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.
CLASSIFICATION WITH DNNS

Training (Development)
- cars
- buses
- trucks
- motorcycles

Inference (Production)
- truck
WHY ARE GPUs GREAT FOR DEEP LEARNING?

<table>
<thead>
<tr>
<th></th>
<th>Neural Networks</th>
<th>GPUs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inherently Parallel</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Matrix Operations</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>FLOPS</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

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

[Lee, Ranganath & Ng, 2007]
CONVOLUTIONAL NEURAL NETWORKS

- Biologically inspired.

- Neuron only connected to a small region of neurons in layer below it called the filter or 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.
CONVOLUTIONAL NET EXAMPLES


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)

Slide credit: Yann Lecun, Facebook & NYU
GPUs - THE PLATFORM FOR MACHINE LEARNING

Image Recognition Challenge

1.2M training images • 1000 object categories

Hosted by

ImageNet

<table>
<thead>
<tr>
<th>2010</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
<th>2014</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>60</td>
<td>110</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

GPU Entries

<table>
<thead>
<tr>
<th>2010</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
<th>2014</th>
</tr>
</thead>
<tbody>
<tr>
<td>0%</td>
<td>5%</td>
<td>10%</td>
<td>15%</td>
<td>20%</td>
</tr>
</tbody>
</table>

Classification Error Rates

<table>
<thead>
<tr>
<th>2010</th>
<th>2011</th>
<th>2012</th>
<th>2013</th>
<th>2014</th>
</tr>
</thead>
<tbody>
<tr>
<td>28%</td>
<td>26%</td>
<td>16%</td>
<td>12%</td>
<td>7%</td>
</tr>
</tbody>
</table>
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"

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
cuDNN  version 2
CuDNN DESIGN GOALS

- **Basic Deep Learning Subroutines**
  - Allow user to write a DNN application without any custom CUDA code

- **Flexible Layout**
  - Handle any data layout

- **Memory - Performance tradeoff**
  - Good performance with minimal memory use, great performance with more memory use
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
- Layout and configuration variations
- Other cuDNN routines have straightforward implementations

| 1 | Minibatch Size |
| 2 | Input feature maps |
| 3 | Image Height |
| 4 | Image Width |
| 5 | Output feature maps |
| 6 | Kernel Height |
| 7 | Kernel Width |
| 8 | Top zero padding |
| 9 | Side zero padding |
| 10 | Vertical stride |
| 11 | Horizontal stride |
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 V2 FLEXIBILITY

Can now specify a strategy the library will use to select the best convolution algorithm:

- PREFER_FASTEST
- NO_WORKSPACE
- SPECIFY_WORKSPACE_LIMIT

...or specify an algorithm directly...

- GEMM
- IMPLICIT_GEMM
- IMPLICIT_PRECOMP_GEMM
- DIRECT
cuDNN V2 NEW FEATURES

Other key new features:

- Support for 3D datasets. Community feedback desired!
- OS X support
- Zero-padding of borders in pooling routines
- Parameter scaling
- Improved support for arbitrary strides
- Support for upcoming Tegra X1 via JIT compilation

See Release Notes for details...
cuDNN V2 API CHANGES

Important - API Has Changed

- Several of the new improvements required changes to the cuDNN API.
- Applications previously using cuDNN V1 are likely to need minor modifications.
- Note Im2Col function is currently exposed public function…but will be removed.

The cuDNN team genuinely appreciates all feedback from the Deep learning community.

The team carefully considers any API change.

cuDNN is still young...API changes expected to become rare in the future.
Using cuDNN
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
Choose a default network, modify one, or create your own

Create your database

Configure your Network

DIGITS Workflow

Create your database

Configure your model

Start training
DIGITS

Visualize DNN performance in real time
Compare networks
Try it today!
developer.nvidia.com/cuDNN