Deep Learning and GPUs

Julie Bernauer
GPU Computing
GPU Computing
CUDA

Framework to Program NVIDIA GPUs

A simple sum of two vectors (arrays) in C

```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

```c
__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

Numerical & Math
RAND, Statistics

Data Struct. & AI
Sort, Scan, Zero Sum

Visual Processing
Image & Video

NVIDIA cuFFT, cuBLAS, cuSPARSE

NVIDIA cuRAND

NVIDIA cuBLAS, cuSPARSE
Deep Neural Networks
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
Deep Neural Networks and GPUs
ACCELERATING INSIGHTS

“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

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

<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012</td>
<td>Google Brain</td>
</tr>
<tr>
<td>2016</td>
<td>AlphaGo</td>
</tr>
</tbody>
</table>

**IMAGENET Accuracy Rate**

- **Traditional CV**
- **Deep Learning**

![Graph showing accuracy rate over years](image)

*At last — a computer program that can beat a champion Go player*
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
IBM

CHAINER
Preferred Networks

THEANO
Université de Montréal

MATCONVNET
University of Oxford

TENSORFLOW
Google

CNTK
Microsoft

TORCH
facebook

CAFFE
Berkeley

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
- Up to 3.5x faster AlexNet training in Caffe than baseline GPU

Caffe Performance

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](https://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
}
Platform
An end-to-end solution
CUDA for deep learning development

<table>
<thead>
<tr>
<th>DEEP LEARNING SDK</th>
</tr>
</thead>
<tbody>
<tr>
<td>DIGITS</td>
</tr>
<tr>
<td>cuDNN</td>
</tr>
<tr>
<td>cuSPARSE</td>
</tr>
<tr>
<td>cuBLAS</td>
</tr>
<tr>
<td>NCCL</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TITAN X</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="TITAN X" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>DEVBOX</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image2" alt="DEVBOX" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>GPU CLOUD</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image3" alt="GPU Cloud" /></td>
</tr>
</tbody>
</table>

- DIGITS
- cuDNN
- cuSPARSE
- cuBLAS
- NCCL
- TITAN X
- DEVBOX
- GPU Cloud
Tesla for hyperscale datacenters

<table>
<thead>
<tr>
<th>HyperScale Suite</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deep Learning SDK</td>
</tr>
<tr>
<td>GPU REST Engine</td>
</tr>
<tr>
<td>GPU Accelerated FFmpeg</td>
</tr>
<tr>
<td>Image Compute Engine</td>
</tr>
<tr>
<td>GPU support in Mesos</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tesla M40</th>
</tr>
</thead>
<tbody>
<tr>
<td>POWERFUL: Fastest Deep Learning Performance</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tesla M4</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOW POWER: Highest Hyperscale Throughput</td>
</tr>
</tbody>
</table>
Jetson for Intelligent machines

**JETSON SDK**

**JETSON TX1**

<table>
<thead>
<tr>
<th>Feature</th>
<th>JETSON TX1</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPU</td>
<td>1 TFLOP/s 256-core Maxwell</td>
</tr>
<tr>
<td>CPU</td>
<td>64-bit ARM A57 CPUs</td>
</tr>
<tr>
<td>Memory</td>
<td>4 GB LPDDR4</td>
</tr>
<tr>
<td>Storage</td>
<td>16 GB eMMC</td>
</tr>
<tr>
<td>Size</td>
<td>50mm x 87mm</td>
</tr>
<tr>
<td>Power</td>
<td>Under 10W</td>
</tr>
</tbody>
</table>

**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 / Nor
- Phase
- Table Lookup
- Threshold

**GEOMETRIC TRANSFORMS**
- Affine Warp
- Warp Perspective
- Flip Image
- Remap
- Staple Image

**FILTERS**
- BoxFilter
- Convolution
- Dilation Filter
- Erosion Filter
- Gaussian Filter
- Gaussian Pyramid
- Laplacian3x3
- Median Filter
- Scharr3x3
- Sobel3x3

**ANALYSIS**
- Histogram
- Histogram Equalization
- Integral Image
- Mean / Std Deviation
- Min / Max / Locations.
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

---

Reduce Training Time from 8 Days to 1 Day

---

**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*
What else can one do with Deep Learning?
Action Recognition
Segmentation

Clement Farabet, Camille Couprie, Laurent Najman and Yann LeCun: Learning Hierarchical Features for Scene Labeling, IEEE Transactions on Pattern Analysis and Machine Intelligence, August, 2013
https://www.youtube.com/watch?v=KkNhdINs13U
Image Captioning

[Diagram showing Recurrent Neural Network and Convolutional Neural Network with a bird perched on a branch of a tree captioned as "a bird perched on a branch of a tree."]

“Automated Image Captioning with ConvNets and Recurrent Nets”

—Andrej Karpathy, Fei-Fei Li
Playing Games

Google’s DeepMind

https://www.youtube.com/watch?v=V1eYniJ0Rnk
Want to try?
Links and resources


Hands-on labs https://nvidia.qwiklab.com/

Question? Email jbernauer@nvidia.com