Deep Learning on GPUs

March 2016
AGENDA

What is Deep Learning?
GPUs and DL
DL in practice
Scaling up DL
What is Deep Learning?
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
Traditional machine perception

Hand crafted feature extractors

Raw data → Feature extraction → Classifier/detector → Result

- SVM,
- shallow neural net,
-...

- HMM,
- shallow neural net,
-...

- Speaker ID,
- speech transcription,
-...

- Topic classification,
- machine translation,
- sentiment analysis...
Deep learning approach

Train:
- Dog
- Cat
- Honey badger

Deploy:
- Dog

Errors:
- Dog
- Cat
- Raccoon

MODEL

- Dog ✔
Artificial neural network

A collection of simple, trainable mathematical units that collectively learn complex functions

Given sufficient training data an artificial neural network can approximate very complex functions mapping raw data to output decisions.
Artificial neurons

Biological neuron

From Stanford cs231n lecture notes

Artificial neuron

\[ y = F(w_1 x_1 + w_2 x_2 + w_3 x_3) \]

\[ F(x) = \max(0, x) \]
Deep neural network (dnn)

Application components:

- Task objective: e.g. Identify face
- Training data: 10-100M images
- Network architecture: ~10 layers, 1B parameters
- Learning algorithm: ~30 Exaflops, ~30 GPU days
Deep learning benefits

- **Robust**
  - No need to design the features ahead of time - features are automatically learned to be optimal for the task at hand
  - Robustness to natural variations in the data is automatically learned

- **Generalizable**
  - The same neural net approach can be used for many different applications and data types

- **Scalable**
  - Performance improves with more data, method is massively parallelizable
Baidu Deep Speech 2
End-to-end Deep Learning for English and Mandarin Speech Recognition

English and Mandarin speech recognition

Transition from English to Mandarin made simpler by end-to-end DL

No feature engineering or Mandarin-specifics required

More accurate than humans

Error rate 3.7% vs. 4% for human tests

http://svail.github.io/mandarin/

http://arxiv.org/abs/1512.02595
AlphaGo
First Computer Program to Beat a Human Go Professional

Training DNNs: 3 weeks, 340 million training steps on 50 GPUs

Play: Asynchronous multi-threaded search

- Simulations on CPUs, policy and value DNNs in parallel on GPUs
- Single machine: 40 search threads, 48 CPUs, and 8 GPUs
- Distributed version: 40 search threads, 1202 CPUs and 176 GPUs

Outcome: Beat both European and World Go champions in best of 5 matches

http://www.nature.com/nature/journal/v529/n7587/full/nature16961.html
http://deepmind.com/alpha-go.html
Deep Learning for Autonomous vehicles
Deep Learning Synthesis

Texture synthesis and transfer using CNNs. Timo Aila et al., NVIDIA Research
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
The Big Bang in Machine Learning

“Google’s AI engine also reflects how the world of computer hardware is changing. (It) depends on machines equipped with GPUs... And it depends on these chips more than the larger tech universe realizes.”

WIRED
GPUs and DL

USE MORE PROCESSORS TO GO FASTER
Deep learning development cycle

Development

Data Scientist

Solver
Network
Dashboard

Select Training Dataset
Select Solver
Design the Neural Network
Train
Profile and Debug

Deployment

Cloud
Feedback

Model

Classification
Detection
Segmentation
Three Kinds of Networks

DNN - all fully connected layers

CNN - some convolutional layers

RNN - recurrent neural network, LSTM
DNN
Key operation is dense $M \times V$

Backpropagation uses dense matrix-matrix multiply starting from softmax scores
**DNN**

**Batching for training and latency insensitive.**

$M \times M$

Batched operation is $M \times M$ - gives re-use of weights.

Without batching, would use each element of Weight matrix once.

Want 10-50 arithmetic operations per memory fetch for modern compute architectures.
CNN

Requires convolution and \( M \times V \)

Filters conserved through plane

Multiply limited - even without batching.

6D Loop

For each output map \( j \)

For each input map \( k \)

For each pixel \( x, y \)

For each kernel element \( u, v \)

\[
B_{xyj} \; +\; A_{(x-u)(y-v)k} \times K_{uvkj}
\]
Other Operations
To finish building a DNN

Pooling

ReLU
(or other non-linear function)

Weight Update

$w_{ij} += \alpha a_j g_i$

These are not limiting factors with appropriate GPU use.

Complex networks have hundreds of millions of weights.
Lots of Parallelism Available in a DNN

- Inputs
- Points of a feature map
- Filters
- Elements within a filter

- Multiplies within layer are independent
- Sums are reductions
- Only layers are dependent
- No data dependent operations

=> can be statically scheduled
TESLA M40
World’s Fastest Accelerator for Deep Learning Training

Reduce Training Time from 13 Days to just 1 Day

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

Note: Caffe benchmark with AlexNet,
CPU server uses 2x E5-2680v3 12 Core 2.5GHz CPU, 128GB System Memory, Ubuntu 14.04
Comparing CPU and GPU - server class

Xeon E5-2698 and Tesla M40

NVIDIA Whitepaper "GPU based deep learning inference: A performance and power analysis."
DL in practice
The Engine of Modern AI

**EDUCATION**

<table>
<thead>
<tr>
<th>TORCH</th>
<th>CAFFE</th>
</tr>
</thead>
<tbody>
<tr>
<td>NYU</td>
<td>Berkeley</td>
</tr>
<tr>
<td>Facebook</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>THEANO</th>
<th>MATCONVNET</th>
</tr>
</thead>
<tbody>
<tr>
<td>Université de Montréal</td>
<td>University of Oxford</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MOCHA.JL</th>
<th>PURINE</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIT</td>
<td>NUS</td>
</tr>
<tr>
<td>Massachusetts Institute of Technology</td>
<td>National University of Singapore</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MINERVA</th>
<th>MXNET*</th>
</tr>
</thead>
<tbody>
<tr>
<td>NYU</td>
<td>NYU</td>
</tr>
<tr>
<td>Microsoft</td>
<td>University of Washington</td>
</tr>
<tr>
<td>Carnegie Mellon University</td>
<td></td>
</tr>
</tbody>
</table>

**BIG SUR**

<table>
<thead>
<tr>
<th>TENSORFLOW</th>
<th>WATSON</th>
<th>CNTK</th>
</tr>
</thead>
<tbody>
<tr>
<td>facebook</td>
<td>Google</td>
<td>IBM</td>
</tr>
<tr>
<td></td>
<td></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>SKYMIN</td>
<td>SCHULTS LABORATORIES</td>
<td>VITRUVIAN</td>
</tr>
</tbody>
</table>

**NVIDIA GPU PLATFORM**

*U. Washington, CMU, Stanford, TuSimple, NYU, Microsoft, U. Alberta, MIT, NYU Shanghai*
# CUDA for Deep Learning Development

## DEEP LEARNING SDK

<table>
<thead>
<tr>
<th>DIGITS</th>
<th>cuDNN</th>
<th>cuSPARSE</th>
<th>cuBLAS</th>
<th>NCCL</th>
</tr>
</thead>
</table>

## TITAN X

![TITAN X Image]

## DEVBOX

![DEVBOX Image]

## GPU CLOUD

- Amazon Web Services
- Microsoft Azure
- IBM SoftLayer
- Alibaba Cloud Computing
cuDNN
Deep Learning Primitives

Accelerating Artificial Intelligence

developer.nvidia.com/cudnn

- GPU-accelerated Deep Learning subroutines
- High performance neural network training
- Accelerates Major Deep Learning frameworks: Caffe, Theano, Torch, TensorFlow
- Up to 3.5x faster AlexNet training in Caffe than baseline GPU

Tiled FFT up to 2x faster than FFT

Millions of Images Trained Per Day

Deep Learning Primitives
CUDA BOOSTS DEEP LEARNING 5X IN 2 YEARS

AlexNet training throughput based on 20 iterations, CPU: 1x E5-2680v3 12 Core 2.5GHz, 128GB System Memory, Ubuntu 14.04
NVIDIA DIGITS
Interactive Deep Learning GPU Training System

Process Data

Configure DNN

Monitor Progress

Visualize Layers

developer.nvidia.com/digits
<table>
<thead>
<tr>
<th>Device</th>
<th>Use Case</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tesla</td>
<td>for Cloud</td>
</tr>
<tr>
<td>Titan X</td>
<td>for PC</td>
</tr>
<tr>
<td>DRIVE PX</td>
<td>for Auto</td>
</tr>
<tr>
<td>Jetson</td>
<td>for Embedded</td>
</tr>
</tbody>
</table>
Scaling DL
Scaling Neural Networks

Data Parallelism

Notes:
Need to sync model across machines.
Largest models do not fit on one GPU.
Requires P-fold larger batch size.
Works across many nodes – parameter server approach – linear speedup.

Adam Coates, Brody Huval, Tao Wang, David J. Wu, Andrew Ng and Bryan Catanzaro
Multiple GPUs

Near linear scaling - data parallel.

Scaling Neural Networks

Model Parallelism

Notes:

- Allows for larger models than fit on one GPU.
- Requires much more frequent communication between GPUs.
- Most commonly used within a node – GPU P2P.
- Effective for the fully connected layers.

Adam Coates, Brody Huval, Tao Wang, David J. Wu, Andrew Ng and Bryan Catanzaro
Scaling Neural Networks

Hyper Parameter Parallelism

Try many alternative neural networks in parallel – on different CPU / GPU / Machines. Probably the most obvious and effective way!
Deep Learning Everywhere

Contact: jbarker@nvidia.com