DEEP LEARNING
Assessing Analytics Project Feasibility and Its Computational Requirements
ABOUT ME
Adam Grzywaczewski - adamg@nvidia.com

- Deep Learning Solution Architect @ NVIDIA - Supporting delivery of AI / Deep Learning solutions

- 10 years experience delivering Machine Learning of all scale (from embedded, mobile to Big Data)

- My past experience:
  - Capgemini: https://goo.gl/MzgGbq
  - Jaguar Land Rover Research: https://goo.gl/ar7LuU
"Just as electricity transformed almost everything 100 years ago, today I actually have a hard time thinking of an industry that I don’t think AI will transform in the next several years, ..." 

Andrew Ng, Founder of Google Brain
ACCELERATED ANALYTICS USE CASES

AUTOMOTIVE
Auto sensors reporting location, problems

COMMUNICATIONS
Location-based advertising

CONSUMER PACKAGED GOODS
Sentiment analysis of what’s hot, problems

FINANCIAL SERVICES
Risk & portfolio analysis
New products

EDUCATION & RESEARCH
Experiment sensor analysis

HIGH TECHNOLOGY / INDUSTRIAL MFG.
Mfg. quality
Warranty analysis

LIFE SCIENCES
Clinical trials

MEDIA/ENTERTAINMENT
Viewers / advertising effectiveness

ON-LINE SERVICES / SOCIAL MEDIA
People & career matching

HEALTH CARE
Patient sensors, monitoring, EHRs

OIL & GAS
Drilling exploration sensor analysis

RETAIL
Consumer sentiment

TRAVEL & TRANSPORTATION
Sensor analysis for optimal traffic flows

TELCO
Smart Meter analysis for network capacity, security

LAW ENFORCEMENT & DEFENSE
Threat analysis - social media monitoring, photo analysis
HOW TO BUILD AI PRODUCTS?

AI enabled analytics

• Overview of factors that make AI projects successful
• Computational requirements of the AI workload
WHAT MAKES AI PROJECTS SUCCESSFUL?
### WHAT TYPE OF A PROBLEM IS IT?

Supervised Learning: the mapping from the data to the labels

<table>
<thead>
<tr>
<th>Data</th>
<th>Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image</td>
<td>Name of objects in the image</td>
</tr>
<tr>
<td>Speech</td>
<td>Text</td>
</tr>
<tr>
<td>Video (e.g. football game)</td>
<td>Event statistics (number of football passes)</td>
</tr>
<tr>
<td>Mortgage application</td>
<td>Mortgage risk</td>
</tr>
<tr>
<td>Text</td>
<td>Speech</td>
</tr>
<tr>
<td>English</td>
<td>French</td>
</tr>
<tr>
<td>Click through data</td>
<td>Content recommendations</td>
</tr>
</tbody>
</table>

Based on a presentation from Andrew Ng
WHAT TYPE OF A PROBLEM IS IT?

Noise vs Structure

Lots of noise, little structure - most probably not

Little noise, complex structure - most probably yes

“If a typical person can do a mental task with less than one second of thought, we can probably automate it using AI either now or in the near future.”

Andrew Ng, Founder of Google Brain
DO YOU HAVE ENOUGH LABELED DATA?
The Achilles heel of deep learning: You need a lot of labeled data.

Without a large dataset, deep learning isn’t likely to succeed.

Labels:

- Getting someone to decide the “right” answer can be hard (think about medical imaging)
- If a dataset requires skilled labor to produce labels, this limits scale / affects the cost

Based on a presentation from Bryan Catanzaro
DO YOU HAVE ENOUGH LABELED DATA?
How much data is enough?

“As of 2016, a rough rule of thumb is that a supervised deep learning algorithm will generally achieve acceptable performance with around 5,000 labeled examples per category, and will match or exceed human performance when trained with a dataset containing at least 10 million labeled examples.”

“Working successfully with datasets smaller than this is an important research area, focusing in particular on how we can take advantage of large quantities of unlabeled examples, with unsupervised or semi-supervised learning.”

Ian Goodfellow, Yoshua Bengio, Aaron Courville

WHAT LEVEL OF ACCURACY DO YOU NEED?

Defining and measuring accuracy

How much accuracy you need? (mortgage risk calculation - high, celebrity portal - low)

Aim for lowest acceptable for the product

What is the measure:

- Accuracy (% correct)
- Coverage (% of examples processed)
- Precision (% of detections that are right)
- Recall (% of objects that are detected)
- Amount of error (for regression problems)

- What protective mechanisms to you need to safeguard the system from unavoidable prediction error?

CAN SOMETHING SIMPLER WORK?

Default Baseline Model

- Build the end to end pipeline ASAP and use a non Deep Learning Baseline Model
- Measure accuracy from day 1
- You need a baseline on which to improve:
  - Simple model that you know very well (linear regression, logistic regression, random forest).
  - Boosted Decision trees are a very good baseline model.
- How does the baseline perform in relation to your target accuracy?
- How does the baseline perform in relation to human accuracy?

COMPUTATIONAL REQUIREMENTS?
**DEEP LEARNING WORKFLOW**

Different stages of the process require different solutions

<table>
<thead>
<tr>
<th>Goal</th>
<th>Build the model</th>
<th>Train the model on real data (hyperparameter tuning)</th>
<th>Train the model in production</th>
<th>Serve the model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Build a promising model</td>
<td>Make the model work with real data and optimise it</td>
<td>Prepare fresh model for your customers</td>
<td>Provide functionality using the model (web, mobile, embedded)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Iteration time</th>
<th>Hours</th>
<th>Days-Weeks</th>
<th>Hours - Weeks</th>
<th>Milliseconds</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th># of machines</th>
<th>1 / team member</th>
<th>10s-100s</th>
<th>0 - 10s - Use case specific</th>
<th>Use case specific</th>
</tr>
</thead>
</table>
Modern deep learning models require computation exceeding the capability of conventional CPUs.

Baidu Deep Speach 2 Model:
- 9400 hours of audio recording for English alone
- Training a single model takes tens of exaFLOPs (1 exaFLOP = $10^{18}$ Floating-point Operations)
- Training a single model would take 3-6 weeks on a single GPU and months on a CPU available at the point of writing
- Systems team (focusing on HPC) and Machine Learning team work hand in hand

ImageNet — Models

COMPUTE

„Outrageously large neural networks“ - size does matter

VGG 19 vs Google LSTM using Sparsely-Gated Mixture-of-Experts layer

Importance of model size

Nuts and Bolts of Applying Deep Learning, Andrew Ng, 2016 - https://youtu.be/F1ka6a13S9I
COMPUTE

„Big Data vs Big Model“

“For any size of the data it’s a good idea to always make the data look small by using a huge model.“

“Human brain has about $10^{14}$ parameters and we only live for about $10^9$ seconds.”

Geoffrey Hinton
NEURAL NETWORK COMPLEXITY IS EXPLODING
To Tackle Increasingly Complex Challenges

2015 - Microsoft ResNet
Superhuman Image Recognition

2016 - Baidu Deep Speech 2
Superhuman Voice Recognition

2017 - Google Neural Machine Translation
Near Human Language Translation
COMPUTE
Not only the GPU

CPU workload:
- Data Loading
- Simulators
- Reinforcement learning environments
- Small models
- It is rational to isolate the CPU intensive workload.

MEMORY

Both GPU memory and host RAM are critical for performance

ImageNet models:

Large scale models:
- Some models are too big for a single GPU and need to be spread across multiple devices
- Lower precision training and trading off memory for compute can address the problem to some extent but will not change the trend for increasing model size

COMMUNICATION BETWEEN THE GPUs

Model parallelism and Data parallelism

• Heavy GPU to GPU traffic
• Techniques like 1 bit SGD address parts of the problem but do not change the trend
COMMUNICATION BETWEEN THE GPUS

Real need for model parallelism

Google translation model
COMMUNICATION BETWEEN THE GPUs

Intra-node performance

AllReduce bandwidth (OMB, size=128MB, in GB/s)
COMMUNICATION BETWEEN THE NODES

Data parallelism and training example movement generate significant traffic.
COMMUNICATION BETWEEN THE NODES

Network interface does matter

Multi-system Workload Scaling vs # of Network Ports Per System

- CNTK
- HPL (124 DGX-1)

Relative Performance

Number of EDR IB 100 Gb ports per System
COMMUNICATION BETWEEN THE NODES

Communication technology does matter

AllReduce bandwidth (OMB, size=128MB, in GB/s)

MPI  Baidu Allreduce  NCCL

CNTK scaling
ResNet50, images/s

Ideal  MPI  NCCL

2 nodes x 4 GPUs (IB EDR, PCI Switch)
4 nodes x 8 GPUs (DGX-1: 4x IB EDR, 4x NVLink)
Neural networks are ‘inefficient learners’. Unlike humans they need a lot of labelled data to train:

- Skype Translator Model - 1.4 billion samples
- Baidu English Speech System - 11,400 hours of audio recording
- Google FaceNet - 260 million faces

- Aim for 200MB/s per GPU for file read access.
- When using scale-out nodes with four or eight GPUs, typical 10 Gbps Ethernet links are not capable of these data rates.
- Architect your systems with adequate caching.
### EXAMPLE CONFIGURATION

Focusing on the training workload

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number of CPUs</strong></td>
<td>2 high end</td>
</tr>
<tr>
<td><strong>Number of GPUs</strong></td>
<td>8 P100 / V100</td>
</tr>
<tr>
<td><strong>System Memory</strong></td>
<td>512-1024 GB</td>
</tr>
<tr>
<td><strong>Scale up</strong></td>
<td>NVLINK</td>
</tr>
<tr>
<td><strong>Scale out</strong></td>
<td>100Gb</td>
</tr>
<tr>
<td><strong>Power / Server (W)</strong></td>
<td>3200</td>
</tr>
<tr>
<td><strong>Servers / Rack</strong></td>
<td>4-10</td>
</tr>
<tr>
<td><strong>Power / Rack (kW)</strong></td>
<td>32</td>
</tr>
</tbody>
</table>
DEEP LEARNING CLUSTER

Reference architecture
NVIDIA ACCELERATED ANALYTICS
GPUs in the Data Center

**ANALYZE**

- Accelerated data analysis, collection, clean-up, correlation and modeling
- Accelerate 10-100x over traditional compute.

**VISUALIZE**

- GPU-accelerated drill downs = real time manipulation of data
- Visualize 100x more data at 40x less infrastructure with subsecond response

**AI-ACCELERATE**

- Use deep learning to exploit the vast amount of data now available
- GPU-powered speed, accuracy, and scale = a new wave of A.I.
GPU-ACCELERATION HAS NO LIMITS

**MapD**
MapD is 55x to 1,000x faster than comparable CPU databases on billion+ row datasets

**Kinetica**
Hardware costs that are $\frac{1}{10}$ that of standard in-memory databases

**BlazeGraph**
200-300x speed-up

**Graphistry**
See 100x more data at millisecond speed

**SQream**
The supercomputing powers of the GPU combined with SQream’s patented technology, results in up to 100 times faster analytics performance on terabyte-petabyte scale data sets
NVIDIA GPU DEEP LEARNING EVERYWHERE, EVERY PLATFORM

TESLA
Servers in every shape and size

DGX Systems — The essential deep learning systems for instant productivity

CLOUD
Everywhere

HEWLETT PACKARD
IBM
Quanta Compute
DELL

Lenovo
CRAY
CISCO

Alibaba.com
Amazon
IBM
Baidu

Google
NVIDIA
Microsoft
Don’t miss the world’s most important event for GPU developers
October 10 - 12, 2017 in Munich