REAL WORLD PROBLEM SIMPLIFICATION USING DEEP LEARNING / AI

Dr. Ettikan Kandasamy Karuppiah
Director, Developers’ Ecosystem

2 November 2017
AI IS EVERYWHERE

“Find where I parked my car”

“Find the bag I just saw in this magazine”

“What movie should I watch next?”
Bringing grandmother closer to family by bridging language barrier

Predicting sick baby’s vitals like heart rate, blood pressure, survival rate

Enabling the blind to “see” their surrounding, read emotions on faces
AI FOR PUBLIC GOOD

Increasing public safety with smart video surveillance at airports & malls

Providing intelligent services in hotels, banks and stores

Separating weeds as it harvests, reduces chemical usage by 90%
HOW A DEEP NEURAL NETWORK SEES

Raw data

Low-level features

Mid-level features

High-level features

Image "Audi A7"
KNOW YOUR PROBLEM WELL

Fig. 1: The first figure shows three car images segmented from a traffic video.

Fig. 8: Three license plates and the corresponding segmented blocks.

Fig. 7: Top row: three car images segmented from the traffic video. Middle row: corresponding saliency maps for license plates. Bottom row: detected license plates.
## AI MOMENTUM

<table>
<thead>
<tr>
<th>TODAY</th>
<th>BY 2020</th>
</tr>
</thead>
<tbody>
<tr>
<td>3000</td>
<td>85%</td>
</tr>
<tr>
<td>AI startups</td>
<td>of all customer service interactions will be powered by AI bots</td>
</tr>
</tbody>
</table>
EVERY INDUSTRY HAS AWOKEN TO AI

Organizations engaged with NVIDIA on Deep Learning

- Higher Ed
- Internet
- Healthcare
- Finance
- Automotive
- Developer Tools
- Government
- Others

19,439

1,549

2014 2016
RISE OF GPU COMPUTING

Original data up to the year 2010 collected and plotted by M. Horowitz, F. Labonte, O. Shacham, K. Olukotun, L. Hammond, and C. Batten New plot and data collected for 2010-2015 by K. Rupp
Add GPUs: Accelerate Data Processing & Analytics
Artificial intelligence is the use of computers to simulate human intelligence.

AI amplifies our cognitive abilities — letting us solve problems where the complexity is too great, the information is incomplete, or the details are too subtle and require expert training.

Learning from data — a computer’s version of life experience — is how AI evolves. GPU computing powers the computation required for deep neural networks to learn to recognize patterns from massive amounts of data.

This new computing model sparked the AI era.
NVIDIA DEEP LEARNING SDK
High Performance GPU-Acceleration for Deep Learning

VISION
- Image Classification
- Object Detection

SPEECH
- Voice Recognition
- Language Translation

BEHAVIOR
- Recommendation Engines
- Sentiment Analysis

DEEP LEARNING FRAMEWORKS
- Caffe
- DL4J
- Mocha.jl
- Keras
- CNTK
- MXNet
- Purine
- Pylearn2
- Theano
- Chainer
- Julia
- Theano

DEEP LEARNING
- cuDNN

MATH LIBRARIES
- cuBLAS
- cuSPARSE
- cuFFT

MULTI-GPU
- NCCL
ANNOUNCING TESLA V100
GIANT LEAP FOR AI & HPC
VOLTA WITH NEW TENSOR CORE

21B xtors | TSMC 12nm FFN | 815mm²
5,120 CUDA cores
7.5 FP64 TFLOPS | 15 FP32 TFLOPS
NEW 120 Tensor TFLOPS
20MB SM RF | 16MB Cache
16GB HBM2 @ 900 GB/s
300 GB/s NVLink
NEW TENSOR CORE

New CUDA TensorOp instructions & data formats

4x4 matrix processing array

\[ D[\text{FP32}] = A[\text{FP16}] \times B[\text{FP16}] + C[\text{FP32}] \]

Optimized for deep learning

- Activation Inputs
- Weights Inputs
- Output Results
AI INFERENCE IS THE NEXT GREAT CHALLENGE

EXPLOSION OF NETWORK DESIGN

Convolution Networks
- PReLU
- ReLU
- BatchNorm
- Concat
- Dropout
- Pooling

Recurrent Networks
- LSTM
- GRU
- Highway
- Projection
- Embedding
- BiDirectional

Generative Adversarial Networks
- 3D-GAN
- Rank GAN
- Conditional GAN
- Coupled GAN
- Speech Enhancement GAN
- Latent space GAN

Reinforcement Learning
- A3C
- DDQN
- Dueling DDQN
MODEL COMPLEXITY IS EXPLODING

2015 — Microsoft ResNet
7 ExaFLOPS
60 Million Parameters

2016 — Baidu Deep Speech 2
20 ExaFLOPS
300 Million Parameters

2017 — Google NMT
105 ExaFLOPS
8.7 Billion Parameters
REVOLUTIONARY AI PERFORMANCE
3X Faster DL Training Performance

Googlenet Training Performance (Speedup Vs K80)

- 8x V100 cuDNN7
- 8x P100 cuDNN6
- 4x M40 cuDNN3
- 1x K80 cuDNN2

Over 80x DL Training Performance in 3 Years

LSTM Training (Neural Machine Translation)

- 2X CPU: 15 Days
- 1X P100: 18 Hours
- 1X V100: 6 Hours

3X Reduction in Time to Train Over P100

Multi-Node Training with NCCL2.0 (ResNet-50)

- 8X P100: 18 Hours
- 8X V100: 7.4 Hours
- 64X V100: 1 Hour

85% Scale-Out Efficiency Scales to 64 GPUs with Microsoft Cognitive Toolkit

Neural Machine Translation Training for 13 Epochs | German –> English, WMT15 subset | CPU = 2x Xeon E5-2699 V4 | V100 performance measured on pre-production hardware.

ResNet50 Training for 90 Epochs with 1.28M images dataset | Using Caffe2 | V100 performance measured on pre-production hardware.
Deep Learning-Inferencing:
TESLA V100 DELIVERS NEEDED RESPONSIVENESS WITH UP TO 99X MORE THROUGHPUT

<table>
<thead>
<tr>
<th>Model</th>
<th>Throughput (Images/Sec)</th>
<th>Latency (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GoogleNet</td>
<td>3,270 i/s</td>
<td>7ms</td>
</tr>
<tr>
<td>ResNet-50</td>
<td>4,647 i/s</td>
<td>7ms</td>
</tr>
<tr>
<td>VGG-16</td>
<td>1,658 i/s</td>
<td>7ms</td>
</tr>
</tbody>
</table>

CPU: Xeon E5-2690 V4
GPU Servers: Dual Xeon E5-2690 v4@2.6GHz with 16GB PCIe GPUs configs as shown
Ubuntu 14.04.5, CUDA 9.0.103, cuDNN 7.0.1.13; NCCL 2.0.4, TensorRT pre-release, data set: ImageNet,
GPU Optimal batch size used to achieve 7ms latency; CPU batch size reduced to 1 if latency exceeds 7ms
HPCG Performance Equivalency
Single GPU Server vs Multiple CPU-Only Servers

To arrive at CPU node equivalence, we use measured benchmark with up to 8 CPU nodes. Then we use linear scaling to scale beyond 8 nodes.

HPCG Benchmark
Exercises computational and data access patterns that closely match a broad set of important HPC applications

VERSION
3

ACCELERATED FEATURES
All

SCALABILITY
Multi-GPU and Multi-Node

MORE INFORMATION
http://www.hpcg-benchmark.org/index.html

CPU Server: Dual Xeon E5-2690 v4@2.6GHz, GPU Servers: same CPU server w/ V100s PCIe
CUDA Version: CUDA 9.0.103; Dataset: 256x256x256 local size
To arrive at CPU node equivalence, we use measured benchmark with up to 8 CPU nodes. Then we use linear scaling to scale beyond 8 nodes.
REINVENTING OUR COMMUNITY
REINVENTING OUR COMMUNITY
Scope of Work

Deep Learning for Damage Estimation based on photo of the damage
Custom development for one of top 5 Insurance Companies

Data Processed to Date*  Total Data Available
1000 Claims  90,000 Claims
4500 images  ~380,000 car images

*Time lag in data migration, data clean-up and tagging images
Neural Network Demo

Visualization of the layers of a neural network and the back propagation process

Visualization of the activations of the convolutions over a car damage image

https://www.galacticar.ai/ -> Galaxy.ai : Artificial Intelligence Driven Sedan Damage Estimator
Software as a Service

Annual fee + fixed fee every time API is pinged

Integration with Insurance Mobile app.

Use cases verified by the industry and insurance clients:
1) Whether client should file a claim or not - Claims triaging
2) Claims estimation from damaged sedan vehicle images
(1) DEEP LEARNING IN FINANCE - TRADING

What we predict

Predict the case when price of stock will have a significant change

Threshold

0.5%

+0.5% and above

-0.5% to +0.5%

less than -0.5%

Current Time

2 pm

Prediction Time Spread

1 hour

Prediction Time

3 pm

Fig 3. Three Classifications of Stock Price Range at a Future Time

Result

+2.48% with low σ

Expected improvement of algorithmic trading strategy performance is 1 bps

Generic Bigdata Operations

Data Acquisition
- Unstructured Data Sources
  - Unstructured Data Collection
- Real-time Data Sources
  - Real-time Data Ingestion
- Structured Data Sources
  - Structured Data Access

Data Preparation
- Data Cleansing
  - Data Quality & Integrity
- Data Harmonization
  - Harmonization & Ontology Mapping
- Correction & Assurance
  - Check for Correction Exception
- Harmonization Terminologies
  - Check for different Representation and usage Ontology

Data Governance
- Scrambling & Protecting Fields/Data
  - Data Anonymity & Data Protection

Data Dissemination
- Mobile Sharing
- Tablet Sharing
- PC Sharing
- Push & Pull Data Platform

Data Visualization
- Social Network Understanding
- Data Reporting & Visualization
- Data Interpretation

Data Analytics
- Data Analytics
- Sentiment Analytics
- Data Statistics
- Network Analytics
- Data Statistics

Published Data Marts/Lake
- Data Models & Schema
- Data Modeling

Data Warehousing/Storage (including Network Data)

Virtualized Platform & Security Management
GPU DATABASES ARE EVEN FASTER

1.1 Billion Taxi Ride Benchmarks

Source: MapD Benchmarks on DGX from internal NVIDIA testing following guidelines of Mark Litwintschik’s blogs: Redshift, 6-node ds2.8xlarge cluster & Spark 2.1, 11 x m3.xlarge cluster w/ HDFS
Silicon Valley-based Blue River Technology has developed a deep learning solution called LettuceBot that rolls through a field photographing 5,000 young plants a minute, using algorithms and machine vision to identify each sprout as lettuce or a weed. The company trained their neural network with GPUs and the Caffe deep learning framework.

Accurate within a quarter inch, the LettuceBot automatically pinpoints weeds, underdeveloped sprouts, and overplanted areas and then applies tiny doses of herbicide to maximize crop production.

Researchers from the Costa Rica Institute of Technology and French Agricultural Research Centre for International Development developed a deep learning algorithm to automatically identify plant specimens that have been pressed, dried and mounted on herbarium sheets.

According to the researchers, this is the first attempt to use deep learning to tackle the difficult taxonomic task of identifying species in natural-history collections.

How does it work?

**Train**
- Whole slide image
- Sample
- Normal
- Tumor
- Training data

**Test**
- Whole slide image
- Overlapping image patches
- Tumor prob. map

**Convolutional Neural Network**

P(tumor)
SAFE AND SMART CITIES IS AN AI PROBLEM

1 billion installed security cameras WW (2020)
30 billion frames per day

Challenging real world conditions
Traditional video analytics not trustworthy

AI achieves super human results
AI driven intelligent video analytics

Image Classification

Accuracy

Human

Deep Learning

Hand-coded CV
<table>
<thead>
<tr>
<th>Detection Trends:</th>
<th>Males</th>
<th>Females</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

NOTE: Objects displayed are subject to a configurable detection duration.
Robust Material Classification with a Tactile Skin Using Deep Learning

Shiv S. Baishya         Berthold Bäuml

Autonomous Learning Robots Lab
Institute of Robotics and Mechatronics
German Aerospace Center (DLR)

shiv.baishya@dlr.de       berthold.baeuml@dlr.de
AI FOR
VISUAL SEARCH
IN MARKET PLACE
# TESLA PLATFORM

Leading Data Center Platform for HPC and AI

## Applications & Services
- Amazon Alexa
- IBM Watson
- Microsoft Cognitive Services
- NVIDIA DIGITS
- Google Cloud Platform Machine Learning Services

## Industry Tools
- Caffe
- mxnet
- Theano
- TensorFlow
- ResNet
- GoogleNet
- AlexNet
- DeepSpeech
- Inception
- BigLSTM

## NVIDIA SDK
- cuDNN
- cuBLAS
- NCCL
- DeepStream

## Deep Learning SDK
- C/C++
- Fortran
- CUDA C/C++
- PGI OpenACC

## Tesla GPU & Systems
- Tesla GPU
- NVLink
- System OEM
- Cloud
DGX STACK
Fully integrated Deep Learning platform

Instant productivity — plug-and-play, supports every AI framework

Performance optimized across the entire stack

Docker/Container Development and Deployment model

Always up-to-date via the cloud

Mixed framework environments — containerized

Direct access to NVIDIA experts
<table>
<thead>
<tr>
<th>DEEP LEARNING NEEDS...</th>
<th>DEEP LEARNING CHALLENGES</th>
<th>NVIDIA DELIVERS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Scientists</td>
<td>Demand far exceeds supply</td>
<td>DIGITS, DLI Training</td>
</tr>
<tr>
<td>Latest Algorithms</td>
<td>Rapidly evolving</td>
<td>DL SDK, GPU-Accelerated Frameworks</td>
</tr>
<tr>
<td>Fast Training</td>
<td>Impossible -&gt; Practical</td>
<td>DGX-1, P100, P40, TITAN X</td>
</tr>
<tr>
<td>Deployment Platform</td>
<td>Must be available everywhere</td>
<td>TensorRT, P40, P4, Jetson, Drive PX</td>
</tr>
</tbody>
</table>
DEEP LEARNING SOFTWARE

NVIDIA DIGITS™
Interactively manage data and train deep learning models for image classification without the need to write code.
Learn more

developer.nvidia.com/deep-learning

Deep Learning Frameworks
Design and train deep learning models using a high-level interface. Choose a deep learning framework that best suits your needs based on your choice of programming language, platform, and target application.
Learn more

NVIDIA Deep Learning SDK
This SDK delivers high-performance multi-GPU acceleration and industry-vetted deep learning algorithms, and is designed for easy drop-in acceleration for deep learning frameworks.
Learn more

developer.nvidia.com/deep-learning
NVIDIA DEEP LEARNING INSTITUTE
Hands-on training for data scientists and software engineers

Training organizations and individuals to solve challenging problems using Deep Learning

On-site workshops and online courses presented by certified experts

Covering complete workflows for proven application use cases

Self-driving cars, recommendation engines, medical image classification, intelligent video analytics and more

Nvidia AI Conference Singapore
23/24th October 2017
THANK YOU