Fuelling the AI Revolution with Gaming

ALISON B LOWNDES
AI DevRel | EMEA
@alisonblowndes
11 days till Xmas!
The day job

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

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

UTILITIES
Smart Meter analysis for network capacity,

LAW ENFORCEMENT & DEFENSE
Threat analysis - social media monitoring, photo analysis
Definitions

ARTIFICIAL INTELLIGENCE
Early artificial intelligence stirs excitement.

MACHINE LEARNING
Machine learning begins to flourish.

DEEP LEARNING
Deep learning breakthroughs drive AI boom.
DNNs

GPUs

Big Data

"the machine equivalent of experience"

2012

Feynman Diagram of DL, © NVIDIA 2016, author Branislav Kisačanin
Artificial Intelligence
GPU Computing

A kernel runs on an SM > grid of blocks > WARP (32 threads)
The kernel is the unit of work {instruction stream with arguments}
HOW GPU ACCELERATION WORKS

Application Code

Compute-Intensive Functions

5% of Code

Rest of Sequential CPU Code

GPU

+ CPU
CUDA 9 Features Revealed: Volta, Cooperative Groups and More

At the 2017 GPU Technology Conference NVIDIA announced CUDA 9, the latest version of CUDA's powerful parallel computing platform and programming model. CUDA 9 is now available as a free download. In this post I’ll provide an overview of the awesome new features of CUDA 9.

- Support for the Volta GPU architecture, including the new Tesla V100 accelerator;
- Cooperative Groups, a new programming model for managing groups of communicating threads;
- A new API (preview feature) for programming Tensor Core matrix multiply and accumulate operations on Tesla V100.
- Faster library routines for linear algebra, image processing, FFTs, and more;
- New algorithms in cusolver and nvGraph;
API toolset for fine-tuning algorithms to get peak performance

Hyperparameter tuning as a service: https://sigopt.com/
CONVOLUTION

Centre element of the kernel is placed over the source pixel. The source pixel is then replaced with a weighted sum of itself and nearby pixels.

http://www.biomachina.org/
CONVOLUTIONAL NEURAL NETWORKS

HOW DOES IT WORK?

Modelled on the Human Brain and Nervous System

TRAINING
Learning a new capability from existing data

Untrained Neural Network Model
Deep Learning Framework

Trained Model
New Capability

TRAINING DATASET
“Suzuki”

Suzuki
HOW DOES IT WORK?

Trained Model
New Capability

INFEERENCE
Applying this capability to new data

NEW DATA
App or Service Featuring Capability

MV Agusta F3

Trained Model
Optimized for Performance
TRAINING VS INFERENCE

TRAINING

Large N

forward

backward

error

“car”

I

?}

Inference

Smaller, varied N

forward

“car”
BIOLOGICAL VS ARTIFICIAL “NEURON”

A cartoon drawing of a biological neuron (left) and its mathematical model (right).

Source: @karpathy U.Stanford course CS231n http://cs231n.github.io
Long short-term memory (LSTM)

Hochreiter (1991) analysed vanishing gradient “LSTM falls out of this almost naturally”

Gates control importance of the corresponding activations

Long time dependencies are preserved until input gate is closed (−) and forget gate is open (O)

Training via backprop unfolded in time

Fig from Vinyals et al, Google April 2015 NIC Generator

Fig from Graves, Schmidhuber et al, Supervised Sequence Labelling with RNNs
ARCHITECTURES

Neural Networks

Perceptron (P)  Feed Forward (FF)  Radial Basis Network (RBF)

Recurrent Neural Network (RNN)  Long / Short Term Memory (LSTM)  Gated Recurrent Unit (GRU)

Auto Encoder (AE)  Variational AE (VAE)  Denoising AE (DAE)  Sparse AE (SAE)

Brain Computer Interfaces
Focused on treatment for
disease and dysfunction
e.g. epilepsy, depression,
Parkinson's but ultimately
to advance human
intelligence by restoring
and extending cognitive
vibrancy.
Why Generative Models

Generative models have a role in many problems.

- Products
  - Super-resolution, Compression, Text-to-speech

- AI
  - Planning, Exploration
  - Intrinsic motivation
  - Model-based RL

- Science
  - Proteomics, Drug Discovery, Astronomy, High-energy physics
Generative Adversarial Networks

T-SNE clustering => the GAN learns the notion of what makes images similar and how to make them close in latent space.

Totally unsupervised the model learns discriminative features of the dataset without ever being told what they are.

Encoder-decoder CNNs learn the mapping from input to output.

Provides the loss function so we don’t have to.

High-res neural inpainting

https://github.com/leehomyc/High-Res-Neural-Inpainting

by Harry Yang, USC
AI DRIVES NEW DISCOVERIES

Using HPC and deep learning, InSilico Medicine created the DeepPharma platform which provides pharmaceutical companies rapid and accurate analysis of massive amounts of data. DeepPharma paves the way for personalized medicine.
WaveNet & PDD (probability density distillation)
AIVA, AI MUSIC COMPOSITION
ROBOTS

SENSE / PERCEIVE

THINK / REASON

ACT
DEEP REINFORCEMENT LEARNING

Sensory Inputs

Inference

Recognition

user task

Perceptron

RNN

Motor PWM

Goal/Reward

MOTION CONTROL

AUTONOMOUS NAVIGATION

Short-term

Long-term
ATLAS, Boston Dynamics
NNAISENSE (IDSIA)
One-shot imitation learning
“AlphaGo’s play makes us feel free, that no move is impossible. Now everyone is trying to play in a style that hasn’t been tried before.”
Zhou Ruiyang, 9 Dan Professional
UNIVERSE

https://universe.openai.com/
THE DREAMS OF SCIENCE FICTION

Photorealistic models
Interactive physics
Collaboration
Early access in September

Photoreal graphics
Physics
AI
Multi-user VR

TRAINED DNN ON NVIDIA JETSON

NVIDIA DGX-1

ISAAC LAB
ISAAC — ACCELERATED LEARNING FOR A WORLD OF INTELLIGENT MACHINES

The Isaac robot simulator, an AI-based software platform, lets developers train robots in highly realistic, physics-based virtual environments and then transfer that knowledge to real-world units. Developers can set up extensive test scenarios using deep learning training, and then simulate them in minutes instead of months.
A massive Deep Learning challenge
Autonomous vehicles will modernize the $10 trillion transportation industry — making our roads safer and our cities more efficient. NVIDIA DRIVE™ PX is a scalable AI car platform that spans the entire range of autonomous driving.

Toyota recently joined some 225 companies around the world that have adopted the NVIDIA DRIVE PX platform for autonomous vehicles. They range from car companies and suppliers, to startups and research organizations.
STATE-OF-THE-ART DRIVERLESS VEHICLES
“PEGASUS”
ROBOTAXI DRIVE PX

320 TOPS CUDA TensorCore | 16x GMSL | 4x 10G | 8x 1G | 16x 100M | Auto-grade | ASIL D
500W | Late Q1 Early Access Partners

Supercomputing Data Center in your Trunk
ANNOUNCING XAVIER DLA
NOW OPEN SOURCE

Command Interface

Tensor Execution Micro-controller

Memory Interface

Input DMA (Activations and Weights)

Unified 512KB Input Buffer
Activations and Weights

Sparse Weight Decompression

Native Winograd Input Transform

MAC Array
2048 Int8 or 1024 Int16 or 1024 FP16

Output Accumulators

Output Postprocess or (Activation Function, Pooling etc.)

Output DMA

Input DMA (Activations and Weights)
Tsvi Achler - Explainable AI
achler@OptimizingMind.com
FLoC 2018

06–19 JULY 2018
OXFORD, UK

FEDERATED LOGIC CONFERENCE 2018

CAV
International Conference on Computer Aided Verification

CSF
IEEE Computer Security Foundations Symposium

FM
International Symposium on Formal Methods

FSCD
International Conference on Formal Structures for Computation and Deduction

ICLP
International Conference on Logic Programming

ICAR
International Joint Conference on Automated Reasoning

ITP
International Conference on Interactive Theorem Provers

LICS
Annual ACM/IEEE Symposium on Logic in Computer Science

SAT
International Conference on Theory and Applications of Satisfiability Testing

FoPSS
School on Foundations of Programming and Software Systems

FLoC Workshops

FLoC Chairs
Daniel Kroening, Marta Kwiatkowska, Moshe Vardi

@FLoC2018
floc2018.org
Machine Learning Security

"addressing where ML works, but may easily be broken"

https://github.com/openai/cleverhans
http://www.mlsecproject.org/
A PLETHORA OF HEALTHCARE STORIES

Molecular Energetics For Drug Discovery
AI for Drug Discovery
Medical Decision Making
Treatment Outcomes
Reducing Cancer Diagnosis Errors by 85%
Predicting Toxicology
Predicting Growth Problems
Image Processing
Gene Mutations
Detect Colon Polyps
Predicting Disease from Medical Records
Enabling Detection of Fatty Acid Liver Disease
DEEP VARIANT

Actual sequencer output: ~1 billion ~100 basepair long DNA reads (30x coverage)

True genome sequence: 3 billion bases in 23 contiguous chunks (chromosomes)

CERN has invested $9B to build Large Hadron Collider (LHC) to discover new particles like Higgs-Boson.

LHC generates huge amounts of data, performing up to 40M proton collisions every second at near speed-of-light. Today with CPUs, only 1% of the sensor data is actually processed, limiting new particle discoveries.

CERN has turned to NVIDIA GPUs for the ability to process all of the sensor data. With GPUs, CERN expects 10x increase in the potential of discovering new particles unknown to mankind.
AI IMPACTING ALL SEGMENTS
New Financial Services Applications

**BANKING**
- Transaction fraud
- Credit risk assessment
- Customer acquisition
- Product cross selling
- Call center analytics

**CAPITAL MARKETS**
- Algorithmic investment
- Large trade execution
- Client recommendations
- Compliance
- Cybersecurity
- Risk management

**INSURANCE**
- Claims automation
- Fraud prevention
- Policy underwriting
- Customer satisfaction
- Annuity offering
DEEP LEARNING ACROSS HPC

UIUC & NCSA: ASTROPHYSICS
5,000X LIGO Signal Processing

U. FLORIDA & UNC: DRUG DISCOVERY
300,000X Molecular Energetics Prediction

SLAC: ASTROPHYSICS
Gravitational Lensing: From Weeks to 10ms

PRINCETON & ITER: PARTICLE PHYSICS
50% Higher Accuracy for Fusion Sustainment

U.S. DOE: CLEAN ENERGY
33% More Accurate Neutrino Detection

U. PITT: DRUG DISCOVERY
35% Higher Accuracy for Protein Scoring
GPU-ACCELERATED APPLICATIONS

www.nvidia.com/gpu-ready-apps
NVIDIA DEEP LEARNING SOFTWARE PLATFORM

GATHER AND LABEL

Gather Data

Rapidly label data, guide training, get insights

Curate data sets

TRAINING

DATA MANAGEMENT

TRAINING DATA

TRAINING

TRAINED NETWORK

CNN
RNN
FC

MODEL ASSESSMENT

DEPLOY WITH TENSORRT

EMBEDDED
Jetson TX

AUTOMOTIVE
Drive PX (XAVIER)

DATA CENTER
Tesla (Pascal, Volta)

NVIDIA DEEP LEARNING SDK
NVIDIA SDK
The Essential Resource for GPU Developers

DEEP LEARNING
Deep Learning SDK
High-performance tools and libraries for deep learning

SELF-DRIVING CARS
NVIDIA DriveWorks™
Deep learning, HD mapping and supercomputing solutions, from ADAS to fully autonomous

VIRTUAL REALITY
NVIDIA VRWorks™
A comprehensive SDK for VR headsets, games and professional applications

GAME DEVELOPMENT
NVIDIA GameWorks™
Advanced simulation and rendering technology for game development

ACCELERATED COMPUTING
NVIDIA ComputeWorks™
Everything scientists and engineers need to build GPU-accelerated applications

DESIGN & VISUALIZATION
NVIDIA DesignWorks™
Tools and technologies to create professional graphics and advanced rendering applications

AUTONOMOUS MACHINES
NVIDIA JetPack™
Powering breakthroughs in autonomous machines, robotics and embedded computing

ADDITIONAL RESOURCES
More resources for GPU Developers

developer.nvidia.com
CUDA 9

# CUDA TOOLKIT 9

## Unleashes Power of Volta

- Optimized for Volta:
  - Tensor Cores
  - Second-Generation NVLink
  - HBM2 Stacked Memory

## Faster Libraries

- GEMM Optimizations for RNNs (cuBLAS)
- >20x Faster Image Processing (NPP)
- FFT Optimizations Across Various Sizes (cuFFT)

## Cooperative Thread Groups

- Flexible Thread Groups
- Efficient Parallel Algorithms
  - Synchronize Across Thread Blocks in a Single GPU or Multi-GPUs

## Developer Tools & Platform Updates

- 1.3x Faster Compiling
- New OS and Compiler Support
- Unified Memory Profiling
- NVLink Visualization
## NVIDIA DEEP LEARNING SDK UPDATE

<table>
<thead>
<tr>
<th>GPU-accelerated DL Primitives</th>
<th>Multi-GPU &amp; Multi-node</th>
<th>High-performance Inference Engine</th>
</tr>
</thead>
<tbody>
<tr>
<td>cuDNN 7</td>
<td>GPU0</td>
<td>TensorFlow model reader</td>
</tr>
<tr>
<td>Faster training</td>
<td>GPU1</td>
<td>Object detection</td>
</tr>
<tr>
<td>Optimizations for RNNs</td>
<td>GPU2</td>
<td>INT8 RNNs support</td>
</tr>
<tr>
<td>Leading frameworks support</td>
<td>GPU3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>NCCL 2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Multi-node distributed training (multiple machines)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Leading frameworks support</td>
<td></td>
</tr>
</tbody>
</table>

**Faster training**
**Optimizations for RNNs**
**Leading frameworks support**

**Multi-node distributed training (multiple machines)**
**Leading frameworks support**

**NCCL 2**

**TensorRT 3**
• C++ with Python API
• builds on the original Caffe designed from the ground up to take full advantage of the NVIDIA GPU platform
• fast and scalable: multi-GPU and multi-node distributed training
• lightweight and portable: designed for mobile and cloud deployment

http://caffe2.ai/
https://github.com/caffe2/caffe2
http://caffe2.ai/docs/tutorials
https://devblogs.nvidia.com/parallelforall/caffe2-deep-learning-framework-facebook/
TensorFlow by Google

Benchmarks & Highlights

**TensorFlow Fold**

- Fastest growing for CPU, GPU & distributed
- Flexible - any computation as a data flow graph
- Google team now implemented NCCL so support for multi-GPU will improve. Already in DGX-1.

**Keras**


www.tensorflow.org
Circa 2000 - Torch7 - 4th (using odd numbers only 1,3,5,7)
Web-scale learning in speech, image and video applications

Maintained by top researchers including
Soumith Chintala - Research Engineer @ Facebook

All the goodness of Torch7 with an intuitive Python frontend that focuses on rapid prototyping, readable code & support for a wide variety of deep learning models.

http://pytorch.org/tutorials/
Tofu: Parallelizing Deep Learning Systems with Auto-tiling

Memonger: Training Deep Nets with Sublinear Memory Cost

MinPy: High Performance System with NumPy Interface

MX NET + Apache

Efficiency Portability Flexibility

MULTI CORE – MULTI GPU – MULTI NODE


https://github.com/dmlc/
https://github.com/NVIDIA/keras
CNTK: enabled the MSFT2016 5.9% wer* “human parity” system in conversational speech https://arxiv.org/abs/1610.05256

https://github.com/Microsoft/CNTK

https://notebooks.azure.com/cntk/libraries/tutorials
WHAT’S NEW IN DIGITS 6?

TENSORFLOW SUPPORT
Train TensorFlow Models Interactively with DIGITS

NEW PRE-TRAINED MODELS
Image Classification: VGG-16, ResNet50
Object Detection: DetectNet

DIGITS 6 Release Candidate available now on Docker Hub for testing and feedback
General availability in September
DIGITS 6
includes a pre-trained U-Net (30% loss)


Video by Olaf Ronneberger, UoFreiberg:
https://lmb.informatik.uni-freiburg.de/people/ronneberger/u-net/u-net-teaser.mp4
PRODUCTION READY DEEP LEARNING

Optimised inference with TensorRT

DOWNLOAD FOR FREE : https://developer.nvidia.com/tensorrt
CPU throughput based on measured inference throughput performance on Broadwell-based Xeon E2690v4 CPU, and doubled to reflect Intel's stated claim that Xeon Scalable Processor will deliver 2x the performance of Broadwell-based Xeon CPUs on Deep Learning Inference.
NVIDIA TensorRT 3

Programmable Inference Accelerator

Compiler for Optimized Neural Networks
Weight & Activation Precision Calibration
Layer & Tensor Fusion
Kernel Auto-Tuning
Multi-Stream Execution

Trained Neural Network
Dynamic Tensor Memory
Multi-Stream Execution
Compiled & Optimized Neural Network
TITAN V
THE MOST POWERFUL PC GPU EVER

5,120 CUDA cores
640 NEW Tensor cores

12GB HBM2 Memory  1.7Gbps

110 Tensor TFLOPS
13.8 TFLOPS fp16 | 6.9 fp32
TESLA V100
The Fastest and Most Productive GPU for AI and HPC

Volta Architecture
Most Productive GPU

Tensor Core
125 Programmable TFLOPS Deep Learning

Improved NVLink & HBM2
Efficient Bandwidth

Volta MPS
Inference Utilization

Improved SIMT Model
New Algorithms
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
NEW TENSOR CORE

New CUDA TensorOp instructions & data formats
4x4 matrix processing array
Optimized for deep learning

Activation Inputs
Weights Inputs
Output Results
THE POWER TO RUN MULTIPLE FRAMEWORKS AT ONCE

Container Images portable across new driver versions

Containerized Applications

- **NVIDIA Docker**
  - TensorFlow Tuned SW
  - CNTK Tuned SW
  - Caffe2 Tuned SW
  - Pytorch Tuned SW

- **CUDA RT**
  - Linux Kernel + CUDA Driver

- **Other Frameworks and Apps**
  - Tuned SW

NVIDIA DGX-1™
Productivity That Follows You From Desk to Data Center to Cloud

Access popular deep learning frameworks, NVIDIA-optimized for maximum performance

DGX containers enable easier experimentation and keep base OS clean

Develop on DGX Station, scale on DGX-1 or the NVIDIA Cloud
ANNOUNCING
NVIDIA GPU CLOUD

GPU-accelerated Cloud Platform Optimized for Deep Learning

Containerized in NV Docker | Optimization across the full stack
Always up-to-date | Fully tested and maintained by NVIDIA | Sign up NOW
THREE STEPS TO DEEP LEARNING WITH NGC

SIGN UP
To get an NGC account, go to:
www.nvidia.com/ngcsignup

DEPLOY IMAGE
On Amazon EC2, choose a P3 instance and deploy the NVIDIA Volta Deep Learning AMI for NGC

PULL CONTAINER
Pick your desired framework (TensorFlow, PyTorch, MXNet, etc.), and pull the container into your instance
ANNOUNCING
NVIDIA SATURNV WITH VOLTA

40 PetaFLOPS Peak FP64 Performance | 660 PetaFLOPS DL FP16 Performance | 660 NVIDIA DGX-1 Server Nodes
GPU DEEP LEARNING IS A NEW COMPUTING MODEL

- DGX-1: Training
- Tesla: Inference in the Datacenter
- Jetson: Inference at the Edge
ANNOUNCING JETSON TX2

JETSON TX2
EMBEDDED AI SUPERCOMPUTER

Advanced AI at the edge
JetPack SDK
< 7.5 watts full module
Up to 2X performance or 2X energy efficiency
DUAL OPERATING MODES

MAX-Q: Maximum Efficiency
Maximum energy efficiency
Up to 2x the energy efficiency of Jetson TX1
Less than 7.5 W

MAX-P: Maximum Performance
Maximum performance
Up to 2x the performance of Jetson TX1
Less than 15 W
JETPACK SDK FOR AI @ THE EDGE

Sample Code

Nsight Developer Tools

Multimedia API

Deep Learning
TensorRT
cuDNN

Computer Vision
VisionWorks
OpenCV

Graphics
Vulkan
OpenGL

Media
libargus
Video API

CUDA, Linux4Tegra, ROS

Jetson Embedded Supercomputer: Advanced GPU, 64-bit CPU, Video CODEC, VIC, ISP
Jetson TX2 Developer Kit

€649/£544 Web or retail
€350/£300 education
Order now US and Europe, shipping now
APAC / other regions starting in April

Jetson TX1 Developer Kit reduced to €549/£459 - Jetson TX1/TX2 Developer kits have same price for education
INDUSTRY ADOPTION

Manufacturing
Agriculture
Construction
Inventory Management
Logistics/Retail

Security
Delivery
Inspection
Autonomous UAV
Social
NVIDIA® Jetson™ Developer Challenge

SUBMISSIONS DUE:

7 1 0 3 0 7 2 7
DAYS HOURS MINUTES SECONDS

ONLINE

Duration 23.10.17-18.02.18

Prize pool $42,789

I want to participate

Join to get important updates and submit entry
AI CITY NEEDS AN EDGE TO CLOUD ARCHITECTURE

- CLOUD
  - 1000s of cameras
  - Resource optimization
  - Public safety

- ON-PREM SERVER
  - 10s-100s of cameras
  - Traffic management
  - Airport security

- CAMERA
  - Parking entrance
  - Law enforcement

- CAMERAS ON
  - PREM SERVER
  - 10s - 100s of cameras

- CLOUD
  - 1000s of cameras
Robotics Teaching Kit with ‘Jet’ - ServoCity

Available to Instructors Now!
developer.nvidia.com/teaching-kits
NVIDIA DEEP LEARNING INSTITUTE

Online self-paced labs and instructor-led workshops on deep learning and accelerated computing

Take self-paced labs at www.nvidia.co.uk/dli

View upcoming workshops and request a workshop onsite at www.nvidia.co.uk/dli

Educators can join the University Ambassador Program to teach DLI courses on campus and access resources. Learn more at www.nvidia.com/dli
Don’t miss the world’s most important event for GPU developers
March 26—29, 2018 in Silicon Valley