REVOLUTIONIZING RETAIL WITH ARTIFICIAL INTELLIGENCE

Scott Brubaker, Paul Hendricks & Alex Sabatier

NVIDIA
# Inception Partners & Retail Ecosystem

<table>
<thead>
<tr>
<th>Vision-Based Applications</th>
<th>Robots, Drones</th>
<th>Digital Signage</th>
</tr>
</thead>
<tbody>
<tr>
<td>focal Systems</td>
<td>fellow robots</td>
<td>Advertima</td>
</tr>
<tr>
<td>MotionLoft</td>
<td>marble</td>
<td>Facemedia</td>
</tr>
<tr>
<td>Aipoly</td>
<td>IFM</td>
<td>Myster-e</td>
</tr>
<tr>
<td>MaxXerience</td>
<td>River Systems</td>
<td>Sensaibe</td>
</tr>
<tr>
<td>deepmagic™</td>
<td>dispatch</td>
<td>Scientific</td>
</tr>
<tr>
<td>TRACX POINT</td>
<td>kiwi</td>
<td>Interfacing</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Visual Search, Tagging</th>
<th>Conversational Commerce</th>
<th>Marketing, Analytics</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>MARKABLE</td>
<td>Infinite Analytics</td>
<td>:) Affectiva</td>
<td>N-tech Lab</td>
</tr>
<tr>
<td>panwell</td>
<td>Kata.ai</td>
<td>Artifacia</td>
<td>Modiface</td>
</tr>
<tr>
<td>GROK STYLE</td>
<td>Streamoid</td>
<td>heuritech</td>
<td></td>
</tr>
<tr>
<td>VUE.AI</td>
<td>shelfai</td>
<td>omniisci</td>
<td></td>
</tr>
<tr>
<td>DigitalBridge</td>
<td>shoppr</td>
<td>Netra</td>
<td></td>
</tr>
<tr>
<td>THE TAKE.AI</td>
<td>Slyce</td>
<td>Qopius</td>
<td></td>
</tr>
<tr>
<td>nFlat epex</td>
<td>iLenze</td>
<td>MARIANA IQ</td>
<td></td>
</tr>
<tr>
<td>huew</td>
<td>Syte</td>
<td></td>
<td></td>
</tr>
<tr>
<td>deepomatic</td>
<td>shoppr</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VISENZE™</td>
<td>Shopcinity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ProductAI</td>
<td>Slyce</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Shopcinity</td>
<td>OMNIOUS</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Physical / In-store**

**Online**
AI FOR RETAIL

SHOPPING EXPERIENCE

SUPPLY CHAIN

CORPORATE HEADQUARTERS
SHOPPING EXPERIENCE: STORE (IVA)

LOSS PREVENTION, SHOPPER TRACKING

FRICTIONLESS COMMERCE

INVENTORY ANALYSIS
# TOP RETAIL IVA USE CASES

## LOSS PREVENTION
- Ticket Switching
- Mis-scanning
- Sweethearting
- Security

## STORE ANALYTICS
- Heat Mapping
- Demographic Analysis
- Shopper/Employee tracking
- Stock Out
- Customer Engagement
- Price Matching
- Pick-up

## AUTONOMOUS SHOPPING
- Autonomous Checkout
- Nano Stores
- Smart Cabinets

$50B in annual shrinkage in US alone

50% of top retailers will implement IVA for store analytics

Autonomous checkout locations to increase 4x annually for next 3 years
SHOPPING EXPERIENCE: ONLINE

RECOMMENDATION ENGINE

AR/VR CONSUMER INTERACTION

IMAGE-BASED SEARCH

STITCH FIX

Olay

nVIDIA
RECOMMENDATION ENGINES ON GPU CLOUD

SONG RECOMMENDATIONS

VIDEO RECOMMENDATIONS

TARGETED RECOMMENDATIONS

Spotify's Top Ten Most Popular Curated Playlists

NETFLIX

yelp
AI IN SUPPLY CHAIN

WAREHOUSE OPTIMIZATION

DYNAMIC SUPPLY CHAIN REAL-TIME RE-ROUTING

FORECASTING AND REPLENISHMENT
AI AT CORPORATE HQ

SINGLE VIEW OF CONSUMER

DEMAND SIGNAL ANALYSIS

AD SPEND OPTIMIZATION

PREDICTIVE ANALYTICS
DATA SCIENCE IN RETAIL

Supply Chain Replenishment
Inventory Management
Price Simulation & Management
Prioritize Promotion - Ad Targeting
Marketing Optimization
Personalized Recommendations
Truck Routing
Online Delivery
THE STORE OF THE FUTURE
Future-Proofed IVA Infrastructure

DL-BASED IVA EDGE USE CASES

- Loss Prevention
- Stock Out Reduction
- Store Analytics
- Security

Server Back of Store
Server (6 x T4s)

In-Store
Cameras
Sensors
Jetson AGX Xavier / Nano
NVIDIA VALUE
Comprehensive Platform for Retail IVA

<table>
<thead>
<tr>
<th>NVIDIA DELIVERS</th>
<th>IVA Inference w/NVIDIA T4 GPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speed Up</td>
<td>27X CPU</td>
</tr>
<tr>
<td>Images/second (1080P)</td>
<td>4400</td>
</tr>
<tr>
<td>Metropolis Platform optimized for IVA</td>
<td>DS Inference SDK</td>
</tr>
<tr>
<td></td>
<td>TensorRT</td>
</tr>
<tr>
<td>GPU accelerated IVA Software Partners</td>
<td>70+</td>
</tr>
<tr>
<td>Deep Learning Education</td>
<td>Developer Blogs + IVA DLI</td>
</tr>
</tbody>
</table>

* Based on ResNet-50

GPU hardware accelerator engines for video decoding and encoding support faster than real-time video processing.

Speedup: 27X Faster
ResNet-50
ART OF THE POSSIBLE
The State of AI in Retail

Paul Hendricks
Solutions Architect
phendricks@nvidia.com
INTRODUCTION

• Paul Hendricks is a Solutions Architect at NVIDIA, helping enterprise customers with their deep learning and AI initiatives

• Paul's background is primarily in retail, and has spent the past 5 years working with many Fortune 500 retail companies to implement data science and AI solutions.

• Prior to joining NVIDIA, Paul worked at Victoria’s Secret as a Data Scientist building models to understand customer propensity to purchase and how to optimize assortment in stores.

• Currently, Paul's research at NVIDIA focuses on intelligent video analytics, machine leaning, recommendation systems, GANs, and reinforcement learning.
INTRODUCTION

• Paul Hendricks is a Solutions Architect at NVIDIA, helping enterprise customers with their deep learning and AI initiatives

• Paul's background is primarily in retail, and has spent the past 5 years working with many Fortune 500 retail companies to implement data science and AI solutions.

• Prior to joining NVIDIA, Paul worked at Victoria’s Secret as a Data Scientist building models to understand customer propensity to purchase and how to optimize assortment in stores.

• Currently, Paul's research at NVIDIA focuses on intelligent video analytics, machine leaning, recommendation systems, GANs, and reinforcement learning.
INTELLIGENT VIDEO ANALYTICS
Image Classification

Problem Background

- Input Data: Images, Videos

- Goal: Given an input, identify the class that input belongs to
Object Detection

Problem Background

• Input Data: Images, Videos

• Goal: Given an input, identify objects and output bounding boxes around the objects and their classes

1. Resize image.
2. Run convolutional network.
3. Non-max suppression.
Object Segmentation (Semantic Segmentation)

Problem Background

- Input Data: Images, Videos
- Goal: Given an input, identify objects and output a mapping of pixels to their respective classes

Figure 1. The Mask R-CNN framework for instance segmentation.
LOSS PREVENTION, STORE ANALYTICS, AND FRICTIONLESS CHECKOUT

https://www.standardcognition.com/
LOCALIZING ALGORITHMS
LOCALIZING ALGORITHMS

Single Stage Detectors

- These algorithms regress the bounding boxes as well as classify the object within that bounding box in a single pass
LOCALIZING ALGORITHMS

Single Stage Detectors

• These algorithms regress the bounding boxes as well as classify the object within that bounding box in a single pass
• Computationally efficient and can be very fast during inference
LOCALIZING ALGORITHMS

Single Stage Detectors

• These algorithms regress the bounding boxes as well as classify the object within that bounding box in a single pass
• Computationally efficient and can be very fast during inference
• Examples: YOLOv3, SSD, RetinaNet, RetinaMask
LOCALIZING ALGORITHMS

Single Stage Detectors
• These algorithms regress the bounding boxes as well as classify the object within that bounding box in a single pass
• Computationally efficient and can be very fast during inference
• Examples: YOLOv3, SSD, RetinaNet, RetinaMask

Two Stage Detectors
• These algorithms generate a number of region proposals which are then passed to a CNN and classified
LOCALIZING ALGORITHMS

Single Stage Detectors
• These algorithms regress the bounding boxes as well as classify the object within that bounding box in a single pass
• Computationally efficient and can be very fast during inference
• Examples: YOLOv3, SSD, RetinaNet, RetinaMask

Two Stage Detectors
• These algorithms generate a number of region proposals which are then passed to a CNN and classified
• Slower during inference since regions must be proposed and then evaluated (often redundant if overlaps)
LOCALIZING ALGORITHMS

Single Stage Detectors

- These algorithms regress the bounding boxes as well as classify the object within that bounding box in a single pass
- Computationally efficient and can be very fast during inference
- Examples: YOLOv3, SSD, RetinaNet, RetinaMask

Two Stage Detectors

- These algorithms generate a number of region proposals which are then passed to a CNN and classified
- Slower during inference since regions must be proposed and then evaluated (often redundant if overlaps)
- Often are more accurate than single stage detectors, especially when trained on semantic segmentations
LOCALIZING ALGORITHMS

Single Stage Detectors

- These algorithms regress the bounding boxes as well as classify the object within that bounding box in a single pass
- Computationally efficient and can be very fast during inference
- Examples: YOLOv3, SSD, RetinaNet, RetinaMask

Two Stage Detectors

- These algorithms generate a number of region proposals which are then passed to a CNN and classified
- Slower during inference since regions must be proposed and then evaluated (often redundant if overlaps)
- Often are more accurate than single stage detectors, especially when trained on semantic segmentations
- Examples: Faster RCNN, Mask RCNN
GETTING STARTED

DLI Courses
• Introduction to Object Detection with TensorFlow – https://courses.nvidia.com/courses/course-v1:DLI+L-AV-04+V1

Papers
• Faster RCNN – https://arxiv.org/pdf/1506.01497
• Mask RCNN – https://arxiv.org/abs/1703.06870
• RetinaNet – https://arxiv.org/abs/1708.02002

Libraries
• DarkNet – https://github.com/pjreddie/darknet
• TensorFlow’s Object Detection API – https://github.com/tensorflow/models/tree/master/research/object_detection
• Facebook’s Mask RCNN Benchmark – https://github.com/facebookresearch/maskrcnn-benchmark

Datasets
• Pascal VOC – http://host.robots.ox.ac.uk/pascal/VOC/
• COCO – http://cocodataset.org/
MACHINE LEARNING
DATA SCIENCE IN RETAIL

- Supply Chain Replenishment
- Inventory Management
- Price Management / Markdown Optimization
- Prioritize Promotion And Ad Targeting
- Marketing Optimization
- Personalized Recommendations
- Truck Routing
- Online Delivery
ML WORKFLOW STIFLES INNOVATION

Wrangle Data

Data Sources → Data Lake

Data Preparation

ETL → Data Preparation

Train

Data Preparation → Train → Evaluate → Predictions

Deploy

Time-consuming, inefficient workflow that wastes data science productivity
DATA SCIENCE WORKFLOW WITH RAPIDS
Open Source, End-to-end GPU-accelerated Workflow Built On CUDA

DATA PREPARATION

GPUs accelerated compute for in-memory data preparation
Simplified implementation using familiar data science tools
Python drop-in Pandas replacement built on CUDA C++. GPU-accelerated Spark (in development)
DATA SCIENCE WORKFLOW WITH RAPIDS
Open Source, End-to-end GPU-accelerated Workflow Built On CUDA

MODEL TRAINING
GPU-acceleration of today’s most popular ML algorithms
XGBoost, Random Forest, Linear Regression, PCA, K-means, k-NN, DBScan, tSVD ...
DATA SCIENCE WORKFLOW WITH RAPIDS
Open Source, End-to-end GPU-accelerated Workflow Built On CUDA

DATA → VISUALIZATION → PREDICTIONS

VISUALIZATION
Effortless exploration of datasets, billions of records in milliseconds
Dynamic interaction with data = faster ML model development
Data visualization ecosystem (Graphistry & OmniSci), integrated with RAPIDS
RAPIDS — OPEN GPU DATA SCIENCE

Software Stack

Data Preparation → Model Training → Visualization

- Data Preparation
- Model Training
- Visualization

Software Stack:
- Python
- RAPIDS
- CUDA
- Apache Arrow
- Dask
- CUDF
- CUML
- CUGraph
- CUDNN
- Deep Learning Frameworks
GETTING STARTED

DLI Courses

• Accelerating Data Science Workflows with RAPIDS – https://courses.nvidia.com/courses/course-v1:DLI+L-DS-01+V1

Resources

RAPIDS GitHub – https://github.com/rapidsai

• cuDF – https://github.com/rapidsai/cudf
• cuML – https://github.com/rapidsai/cuml
• cuGraph – https://github.com/rapidsai/cugraph
• XGBoost – https://github.com/rapidsai/xgboost
• Dask cuDF – https://github.com/rapidsai/dask-cudf
• Dask cuML – https://github.com/rapidsai/dask-cuml
• Dask XGBoost – https://github.com/rapidsai/dask-xgboost
• Notebooks – https://github.com/rapidsai/notebooks
• Notebooks Extended– https://github.com/rapidsai/notebooks-extended
NVIDIA HARDWARE
TESLA V100
TENSOR CORE GPU

World’s Most Advanced
Data Center GPU

5,120 CUDA cores
640 Tensor cores
7.8 FP64 TFLOPS | 15.7 FP32 TFLOPS
| 125 Tensor TFLOPS
20MB SM RF | 16MB Cache
32 GB HBM2 @ 900GB/s | 300GB/s NVLink
TENSOR CORE BUILT FOR AI
Delivering 125 TFLOPS of DL Performance

MATRIX DATA OPTIMIZATION:
Dense Matrix of Tensor Compute

TENSOR-OP CONVERSION:
FP32 to Tensor Op Data for Frameworks

VOLTA TENSOR CORE
4x4 matrix processing array
\[ D[\text{FP32}] = A[\text{FP16}] \times B[\text{FP16}] + C[\text{FP32}] \]
Optimized For Deep Learning

VOLTA-OPTIMIZED cuDNN

ALL MAJOR FRAMEWORKS

Caffe2
mxnet
PyTorch
TensorFlow
NVIDIA DGX
AI Supercomputer-in-a-Box

1000 TFLOPS | 8x Tesla V100 32GB | NVLink Hybrid Cube Mesh
2x Xeon | 8 TB RAID 0 | Quad IB 100Gbps, Dual 10GbE | 3U — 3200W
NVIDIA DGX-2
THE WORLD’S MOST POWERFUL DEEP LEARNING SYSTEM
FOR THE MOST COMPLEX DEEP LEARNING CHALLENGES

- First 2 PFLOPS System
- 16 V100 32GB GPUs Fully Interconnected
- NVSwitch: 2.4 TB/s bisection bandwidth
- 24X GPU-GPU Bandwidth
- 0.5 TB of Unified GPU Memory
- 10X Deep Learning Performance
TESLA T4
WORLD’S MOST ADVANCED SCALE-OUT GPU

2,560 CUDA Cores
320 Turing Tensor Cores
65 FP16 TFLOPS | 130 INT8 TOPS | 260 INT4 TOPS
16GB | 320GB/s
70 W
NEW TURING TENSOR CORE

MULTI-PRECISION FOR AI INFERENCE & ENTRY LEVEL TRAINING
65 TFLOPS FP16  |  130 TeraOPS INT8  |  260 TeraOPS INT4
WORLD’S MOST PERFORMANT INFERENCE PLATFORM
Up To 36X Faster Than CPUs | Accelerates All AI Workloads

Speedup: 21X faster
DeepSpeech 2

Speedup: 27X faster
ResNet-50 (7ms latency limit)

Speedup: 36X faster
GNMT

For all three graphs:
Dual-Socket Xeon Gold 6140 @ 3.6GHz with single GPU as shown
18.11-py3 | TensorRT 5.0 | CPU FP32, P4 & T4: INT8 | Batch Size = 128
WORLD’S FASTEST INFEERENCE PERFORMANCE
NVIDIA GPUs Set New Performance Records

Throughput

<table>
<thead>
<tr>
<th></th>
<th>ResNet-50</th>
<th>GoogleNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>NVIDIA T4</td>
<td>4,018</td>
<td>5,760</td>
</tr>
<tr>
<td>NVIDIA V100</td>
<td>6,359</td>
<td>12,300</td>
</tr>
</tbody>
</table>

Latency

<table>
<thead>
<tr>
<th></th>
<th>ResNet-50</th>
<th>GoogleNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>NVIDIA T4</td>
<td>0.65</td>
<td>1.10</td>
</tr>
<tr>
<td>NVIDIA V100</td>
<td>0.60</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Energy Efficiency

<table>
<thead>
<tr>
<th></th>
<th>ResNet-50</th>
<th>GoogleNet</th>
</tr>
</thead>
<tbody>
<tr>
<td>NVIDIA T4</td>
<td>60</td>
<td>85</td>
</tr>
<tr>
<td>NVIDIA V100</td>
<td>27</td>
<td>53</td>
</tr>
</tbody>
</table>
THE JETSON FAMILY

JETSON TX1
7 - 15W
1 TFOPS (FP16)
50mm x 87mm

JETSON TX2
7 - 15W
1.3 TOPS (FP16)
50mm x 87mm

JETSON AGX XAVIER
10 - 30W
10 TFLOPS (FP16) | 32 TOPS (INT8)
100mm x 87mm

Multiple devices • Unified software

AI at the edge
UAVs • AI subsystems • AI Cameras

Fully autonomous machines
Factory automation • Logistics • Delivery robots
NVIDIA SOFTWARE
CHALLENGES WITH DEEP LEARNING

Current DIY deep learning environments are **complex** and **time consuming** to build, test and maintain.

Requires high level of **expertise** to manage driver, library, framework dependencies.

Development of frameworks by the community is moving **very fast**.
NVIDIA GPU CLOUD
Deep Learning Everywhere, For Everyone

Innovate in minutes, not weeks
Removes all the DIY complexity of deep learning software integration

Always up to date
Monthly updates by NVIDIA to ensure maximum performance

Deep learning across platforms
Containers run locally on DGX Systems and TITAN PCs, or on cloud service provider GPU instances

NVIDIA GPU Cloud integrates GPU-optimized deep learning frameworks, runtimes, libraries, and OS into a ready-to-run container, available at no charge
COMMON SOFTWARE STACK ACROSS DGX FAMILY

- Single, unified stack for deep learning frameworks
- Predictable execution across platforms
- Pervasive reach
**TENSORRT DEPLOYMENT WORKFLOW**

**Step 1: Optimize trained model**

1. Import Model
2. TensorRT Optimizer
3. Serialize Engine

Trained Neural Network → TensorRT Optimizer → Optimized Plans

**Step 2: Deploy optimized plans with runtime**

1. De-serialize Engine
2. Deploy Runtime

Optimized Plans → TensorRT Runtime Engine → Deploy Runtime

Plan 1, Plan 2, Plan 3 → Optimized Plans → De-serialize Engine → TensorRT Runtime Engine → Deploy Runtime

Data center, Automotive, Embedded
NVIDIA TENSORRT
From Every Framework, Optimized For Each Target Platform
TensorRT 5 & TensorRT Inference Server

Turing Support • Optimizations & APIs • Inference Server

World’s Most Advanced Inference Accelerator
Up to 40x faster perf. on Turing Tensor Cores

New optimizations & flexible INT8 APIs
New INT8 workflows, Win & CentOS support

TensorRT inference server
Maximize GPU utilization, run multiple models on a node

Free download to members of NVIDIA Developer Program soon at developer.nvidia.com/tensorrt
TensorRT Inference Server

Containerized Microservice for Data Center Inference

Multiple models scalable across GPUs

Supports all popular AI frameworks

Seamless integration into DevOps deployments leveraging Docker and Kubernetes

Ready-to-run container, free from the NGC container registry
TRANSFER LEARNING TOOLKIT
End to End NVIDIA Deep Learning Workflow

Pre-Trained model access from NGC * Training & adaptation * Applications ready to integrate with DeepStream

Accelerate time to market and save on compute resources!
DEEPSTREAM
NVIDIA DEEPSTREAM

Zero Memory Copies