“AI BEYOND THE BUZZWORD: DO IT WELL OR DO IT TWICE”

Walter Riviera
AI – Technical Solution Specialist
EMEA
NOTICES AND DISCLAIMERS

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration.

No product or component can be absolutely secure.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. For more complete information about performance and benchmark results, visit http://www.intel.com/benchmarks.

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit http://www.intel.com/benchmarks.

Intel® Advanced Vector Extensions (Intel® AVX)* provides higher throughput to certain processor operations. Due to varying processor power characteristics, utilizing AVX instructions may cause a) some parts to operate at less than the rated frequency and b) some parts with Intel® Turbo Boost Technology 2.0 to not achieve any or maximum turbo frequencies. Performance varies depending on hardware, software, and system configuration and you can learn more at http://www.intel.com/go/turbo.

Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

Cost reduction scenarios described are intended as examples of how a given Intel-based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction.

Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.

Arria, Celeron, Intel, the Intel logo, Intel Atom, Intel Core, Intel Nervana, Intel Optane, Intel Xeon, Iris, Movidius, OpenVINO, Pentium, Stratix and the Stratix logo and are trademarks of Intel Corporation in the U.S. and/or other countries.

*Other names and brands may be claimed as property of others.
WHAT IS AI?

AI

MACHINE LEARNING

DEEP LEARNING
MACHINE LEARNING VS DEEP LEARNING?

MACHINE LEARNING
How do you engineer the best features?

DEEP LEARNING
Find the best parameters

\[ N \times M \]

\( (f_1, f_2, ..., f_K) \)
Roundness of face
Dist between eyes
Nose width
Eye socket depth
Cheek bone structure
Jaw line length
...etc.

CLASSIFIER ALGORITHM
SVM
Random Forest
Naïve Bayes
Decision Trees
Logistic Regression
Ensemble methods

NEURAL NETWORK

Walter

Walter
HOW DOES DEEP LEARNING WORK?

CHOOSE THE BEST AI APPROACH FOR YOUR CHALLENGE

DEEP LEARNING
- Regression
- Classification
- Clustering
- Decision Trees
- Data Generation
- Image Processing
- Speech Processing
- Natural Language Processing
- Recommender Systems
- Adversarial Networks
- Reinforcement Learning

IMAGE RECOGNITION

TRAINING
- Human
- Bicycle
- Strawberry
- Lots of labeled data!
- Forward
- Backward
- Error
- Model weights

INFERENCE
- Bicycle
- Strawberry
- "Bicycle"
- Forward
- "Bicycle"?

© 2019 Intel Corporation
DEPLOY AI ANYWHERE
WITH UNPRECEDENTED HARDWARE CHOICE

Visit: www.intel.ai/technology

AUTOMATED DRIVING
DEDICATED MEDIA/VISION
FLEXIBLE ACCELERATION
DEDICATED DL INFEERENCE
DEDICATED DL TRAINING
GRAPHICS, MEDIA & ANALYTICS

All products, computer systems, dates, and figures are preliminary based on current expectations, and are subject to change without notice.
© 2019 Intel Corporation
BUSINESS IMPERATIVE

THE AI JOURNEYS

INTEL® AI TOMORROW
50 million customer journeys,
2,000 services per day,
1300 km of track,
82 Stations

- On time arrivals UP by 9%,
- Currently @ 92% of trains on time
- Increased automation: additional 6,000 short term schedules
- Improved station planning
- Over 50,000 schedule adjustments
- Time saved in all sectors: Long distance, Express airport services, Commuter, Freight

# DATA PROBLEM

<table>
<thead>
<tr>
<th>INFO 1</th>
<th>INFO 2</th>
<th>INFO 3</th>
<th>Value to Predict</th>
</tr>
</thead>
<tbody>
<tr>
<td>City/town</td>
<td>District</td>
<td>County</td>
<td>Postcode</td>
</tr>
</tbody>
</table>
# DATA PROBLEM

<table>
<thead>
<tr>
<th>INFO 1</th>
<th>INFO 2</th>
<th>INFO 3</th>
<th>Value to Predict</th>
</tr>
</thead>
<tbody>
<tr>
<td>City/town</td>
<td>District</td>
<td>County</td>
<td>Postcode</td>
</tr>
<tr>
<td>City/town</td>
<td>District</td>
<td>Old/new</td>
<td>Price</td>
</tr>
</tbody>
</table>
## DATA PROBLEM

<table>
<thead>
<tr>
<th>INFO 1</th>
<th>INFO 2</th>
<th>INFO 3</th>
<th>Value to Predict</th>
</tr>
</thead>
<tbody>
<tr>
<td>City/town</td>
<td>District</td>
<td>County</td>
<td>Postcode</td>
</tr>
<tr>
<td>City/town</td>
<td>District</td>
<td>Old/new</td>
<td>Price</td>
</tr>
<tr>
<td>Postcode</td>
<td>Property type</td>
<td>Old/new</td>
<td>Price</td>
</tr>
</tbody>
</table>

**DATA DRIVES DECISIONS**
FRAUD DETECTION

China UnionPay

**Client:** China UnionPay*, which specializes in banking services and payment systems. It is the 3rd largest payment network in the world.

**Challenge:** No single ML or DL algorithm delivers a high-enough accuracy when trying to detect online fraudulent transactions using fake/clone cards through POS, ATM, or online payments. Real time detection is required that can service hundreds of millions of transactions per day.

**Solution:** Intel collaborated with UnionPay* on a GBDT→GRU→RF “sandwich” architecture. The entire solution was based on the Intel® Xeon® Scalable processor, and each layer used Intel-optimized software: GBDT (Apache Spark with BigDL), GRU (TensorFlow* optimized for Intel architecture), and RF (Intel® Python with DAAL).

1 GBDT = “Gradient Boosting Decision Tree,” GRU = “Gated Recurrent Unit.” RF = “Random Forest.”

RESULT

“The new sandwich-structured fraud detection model has performed up to expectations in various assessments by the National Engineering Laboratory for E-commerce and E-payment and ZhongAn Technology.”

---

* Intel® MKL-DNN Intel® Python with DAAL

1 Intel Confidential – Internal Use Only
Fraud detection – Challenges

**Data science challenges:**
- DT (Decision Tree), RF (Random Forest), GBDT not effective enough.
- RNNs not autonomous enough

→ Combined approach: **the “Sandwich” architecture**

**IT challenges:**
- Leverage the existing Data infrastructure with Hadoop storage system
- Enable Tensorflow

→ Intel BigDL
Fraud detection

Vectors for single transactions

Select corresponding model

n - dimensional

WB Structure 1
GRU TS=E_1

WB Structure 2
GRU TS=E_2

WB Structure k
GRU TS=E_k

Transform

n_{op} - dimensional

Optimized vectors for each transaction

Random Forest
Case Study: SubtlePET – Enabling faster PET Scans

SubtlePET enhances the quality of the DICOM images produced by PET scanners using faster scan protocols to maintain clinically-equivalent results.

Customer:
SubtlePET

Challenge:
SubtlePET, for some customers, has to be deployed to computing devices that are already in the hospital. In most cases, this is an Intel® CPU. Improving inference speed on existing clinical infrastructure is paramount to accelerating the adoption of SubtlePET into the clinical workflow.

Solution:
SubtlePET incorporated the Intel® OpenVINO toolkit into their C++ solution which improved image quality and enabled faster PET scans by increasing signal-to-noise ratio. The integration of Intel® distribution of OpenVINO™ toolkit increased the inference speed of their typical PET/CT whole body by 2.1x on Intel® Xeon E5 processor.

RESULTS

2.1X FASTER
Inference performance

Optimization Notice: Intel’s compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice. Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit: http://www.intel.com/performance

Other Marketing Assets: https://builders.intel.com/ai/blog/subtle-medical-subtlepet-openvino-toolkit

*Other names and brands may be claimed as the property of others.

Configuration: AWS p3.2xlarge instance, Intel® Xeon® E5-2686 v4 @ 2.30 GHz, 61 GB RAM, Intel® OpenVINO™ Toolkit.
Intel® Distribution of OpenVINO™ toolkit

**Model Optimizer**
- What it is: Preparation step -> imports trained models
- Why important: Optimizes for performance/space with conservative topology transformations; biggest boost is from conversion to data types matching hardware.

**Inference Engine**
- What it is: High-level inference API
- Why important: Interface is implemented as dynamically loaded plugins for each hardware type. Delivers best performance for each type without requiring users to implement and maintain multiple code pathways.

---

<table>
<thead>
<tr>
<th>Model Optimizer</th>
<th>Inference Engine</th>
</tr>
</thead>
<tbody>
<tr>
<td>▪ What it is: Preparation step -&gt; imports trained models</td>
<td>▪ What it is: High-level inference API</td>
</tr>
<tr>
<td>▪ Why important: Optimizes for performance/space with conservative topology transformations; biggest boost is from conversion to data types matching hardware.</td>
<td>▪ Why important: Interface is implemented as dynamically loaded plugins for each hardware type. Delivers best performance for each type without requiring users to implement and maintain multiple code pathways.</td>
</tr>
</tbody>
</table>

**Training Model**
- Caffe
- TensorFlow*
- ONNX
- MxNet*
- KALDI

**Intermediate Representation (IR)**
- data

**Convert & Optimize**
- IR = Intermediate Representation format

**Load, infer**

---

**Model Optimizer**
- ▪ What it is: Preparation step -> imports trained models
- ▪ Why important: Optimizes for performance/space with conservative topology transformations; biggest boost is from conversion to data types matching hardware.

**Inference Engine**
- ▪ What it is: High-level inference API
- ▪ Why important: Interface is implemented as dynamically loaded plugins for each hardware type. Delivers best performance for each type without requiring users to implement and maintain multiple code pathways.

---

**CPU Plugin**
- Extendibility C++

**GPU Plugin**
- Extendibility OpenCL™

**FPGA Plugin**
- Extendibility OpenCL/TBD

**Myriad Plugin**
- Extendibility TBD

---

GPU = Intel CPU with integrated graphics processing unit/Intel® Processor Graphics
OpenCL and the OpenCL logo are trademarks of Apple Inc. used by permission by Khronos
*Other names and brands may be claimed as the property of others.
The “DeepAd” Project with Cloudian* & Dentsu*

Targeted Advertisement

**Client:** The “DeepAd project” comprised primarily of Cloudian*, Dentsu*, and Intel testing innovative digital billboard solutions with dynamic content.

**Challenge:** Utilizing machine learning and deep learning to increase the effectiveness of OOH (out of home) signage along busy freeways for targeted advertising, detecting and tracking different type of automobiles and displaying content based on the recognized car.

**Solution:** Intel® Xeon® processor E5 family servers used to train a models consisting of over 5000 vehicles, with Intel® NUCs in the field which delivered the speed needed to help accurately detect and track cars – all within less than a second!


"Multi-core Intel® Xeon® Processor E5 family [...] were critical to enabling the speed at which real-time detection, control, and decision making were possible [and the] the Intel NUC kept the technology footprint small in the field…"

*Other names and brands may be claimed as the property of others. Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate.
DEEP-AD - Challenges

**IT challenges:**
- Build an end-to-end solution Datacentre - Edge
- Fast decision making happening at the Edge

→ Intel portfolio of AI solutions (Widest!)
Convergence Today on Intel® Xeon® Processor Based Supercomputing Infrastructure

Better Theory → Guided Data
Science via consistency

Better Theory → Guided Data
Science via consistency

HPC ←→ DA ←→ AI

Equations of motion:
Differential equations and their solutions:
\[ \frac{d^2 x}{dt^2} = \frac{F_x}{m_{AK}} = 0 \]
\[ \frac{d^2 y}{dt^2} = \frac{F_y}{m_{AK}} = 0 \]

Displacement of Analytical Models

Training Data augmenting Real World Data

Data Prep, ETL, Dimension Redux

Data Compression via Learning

Natural Events/Social Media

A World of analytical models
HPC
Modeling & Simulation

A World of Data driven Models
AI

A WORLD OF DATA ANALYTICS
DA

Feature Vectors
for Training

Feature Vectors
for Training
**HPC POC: High Energy Physics Simulation**

**CERN**

*Joint collaboration with Intel and SURFsara and TACC*

---

**Customer:** CERN, the European Organization for Nuclear Research, which operates the Large Hadron Collider (LHC), the world’s largest and most powerful particle accelerator

**Challenge:** CERN currently uses Monte Carlo simulations for complex physics and geometry modeling, which is a heavy computational load that consumes up to >50% of the Worldwide LHC (Large Hadron Collider) Computing Grid (WLCG) power for electron shower simulations.

**Solution:** Distributed training using 128 nodes of the TACC Stampede 2 cluster (Intel® Xeon® Platinum 8160 processor, Intel® OPA) and a 3D Generative Adversarial Network (3D GAN). Performance was first optimized on a single node using TensorFlow® optimized with Intel® MKL-DNN, using 4 workers/node and an optimized number of convolutional filters.

---

**RESULT**

94% scaling efficiency up to 128 nodes, with a significant reduction in training time per epoch for 3D GANs

---

*Other names and brands may be claimed as the property of others.


Performance results are based on testing as of (05/17/2018) and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure.

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSumark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit [http://www.intel.com/performance](http://www.intel.com/performance).

---

*Customer: CERN, the European Organization for Nuclear Research, which operates the Large Hadron Collider (LHC), the world’s largest and most powerful particle accelerator

Challenge: CERN currently uses Monte Carlo simulations for complex physics and geometry modeling, which is a heavy computational load that consumes up to >50% of the Worldwide LHC (Large Hadron Collider) Computing Grid (WLCG) power for electron shower simulations.

Solution: Distributed training using 128 nodes of the TACC Stampede 2 cluster (Intel® Xeon® Platinum 8160 processor, Intel® OPA) and a 3D Generative Adversarial Network (3D GAN). Performance was first optimized on a single node using TensorFlow® optimized with Intel® MKL-DNN, using 4 workers/node and an optimized number of convolutional filters.

94% scaling efficiency up to 128 nodes, with a significant reduction in training time per epoch for 3D GANs

---

*Other names and brands may be claimed as the property of others.


Performance results are based on testing as of (05/17/2018) and may not reflect all publicly available security updates. See configuration disclosure for details. No product can be absolutely secure.

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSumark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products. For more complete information visit [http://www.intel.com/performance](http://www.intel.com/performance).
Multi-Node Training Performance & Accuracy

Distributed training using data parallelism

94% Scaling efficiency up to 128 nodes

High Energy Physics: 3D GANs Training Speedup Performance
Intel 2S Xeon(R) on Stampede2/TACC, OPA Fabric
TensorFlow 1.9+MKL-DNN+horovod, Intel MPI, Core Aff. BKMs, 4 Workers/Node

2S Xeon 8160: Secs/Epoch Speedup
Ideal Scaling Efficiency
128-Node Perf: 148 Secs/Epoch

Ratio of Ecal and Ep
High Energy Physics Simulation - CERN

Paper: https://doi.org/10.1007/978-3-030-02465-9_35

Code: https://github.com/svalleco/3Dgan

Distributed Model: https://github.com/svalleco/3Dgan/blob/svalleco/sc18/keras/

The training script is EcalEnergyTrain_hvd.py

Architecture details: EcalEnergyGan.py

Training dataset: https://cernbox.cern.ch/index.php/s/CczhzHSwLrVWD4p
1. **Know your goal:** Data drives decision

2. **Max ROI:** turn your HPC/BigData facility into an AI capable infrastructure

3. **Intel AI Portfolio:** 1 solution does not fit all needs. Find your best!
WHAT ABOUT TOMORROW?

Input Size

Model Size

Today

Low

High

Tomorrow

Low

High

<1MB

ImageNet

224x224

>1TB

Billions of Parameters

2000x2000x2000

BIGGER DATA

BIGGER MODELS

BIGGER COMPUTE

Source: Siemens Healthineers; https://www.youtube.com/watch?v=o8NIN8dJqdc
Source: https://idealmagazine.co.uk/the-ideal-guide-how-to-order-a-beer-pizzas-taxi-coffee-in-15-languages/

*Other names and brands may be claimed as the property of others
Highly-efficient multi-model inferencing for cloud, data center and intense appliances
To quickly process vast, sparse, or complex data for large models within a power budget, AI hardware must deliver a critical balance of:

- **VECTOR**
- **SCALAR**
- **SPATIAL**
- **MATRIX**
- **COMPUTE**
- **COMMUNICATION**
- **MEMORY**
DIVERSE WORKLOADS REQUIRE DIVERSE ARCHITECTURES

The future is a diverse mix of scalar, vector, matrix, and spatial architectures deployed in CPU, GPU, AI, FPGA and other accelerators.

PROGRAMMING CHALLENGE

Diverse set of data-centric hardware
No common programming languages or API
Inconsistent tool support across platform
Each platform require software investment
INTEL’S ONE API
CORE CONCEPT

One API is a project to deliver a unified programming model to simplify development across diverse architectures

Common developer experience across Scalar, Vector, Matrix and Spatial (SVMS) architecture

Unified and simplified language and libraries for expressing parallelism

Uncompromised native high-level language performance

Support for CPU, GPU, AI and FPGA

Based on industry standards and open specifications

One API Optimized Apps

One API Optimized Middleware / Frameworks

One API Language & Libraries

CPU

GPU

AI

FPGA

SCALAR

VECTOR

MATRIX

SPATIAL

Optimization Notice
Copyright © 2019, Intel Corporation. All rights reserved.
*Other names and brands may be claimed as the property of others.