DELL EMC

Whitepaper

Retail Analytics with Malong RetailAI® on DELL EMC PowerEdge servers

Revision: **1.1**
Issue Date: **10/14/2019**

Abstract

This blog evaluates the performance and efficiency of running the Malong RetailAI® software stack on Dell EMC PowerEdge R7425 server for retail analytics. The objective is to show how the stack can deliver high throughput & low latency inferencing performance on NVIDIA's AI software platform powered by NVIDIA Tensor Core T4 GPUs.

Revisions

Date	Description
02 October 2019	Initial release

Acknowledgements

This paper was produced by the following people:

Name	Role
Bhavesh Patel	Server Advanced Engineering, Dell EMC
Matt Scott	CTO, Malong
Hao Wei	VP of Engineering, Malong

Overview of Deep Learning

Deep learning consists of two phases: Training and inference. As illustrated in Figure 1, training involves learning a neural network model from a given training dataset over a certain number of training iterations and loss function [1]. The output of this phase, the learned model, is then used in the inference phase to speculate on new data.

The major difference between training and inference is training employs *forward propagation* and *backward propagation* (two classes of the deep learning process) whereas inference mostly consists of forward propagation [2]. To generate models with good accuracy, the training phase involves several training iterations and substantial training data samples, thus requiring many-core CPUs or GPUs to accelerate performance.

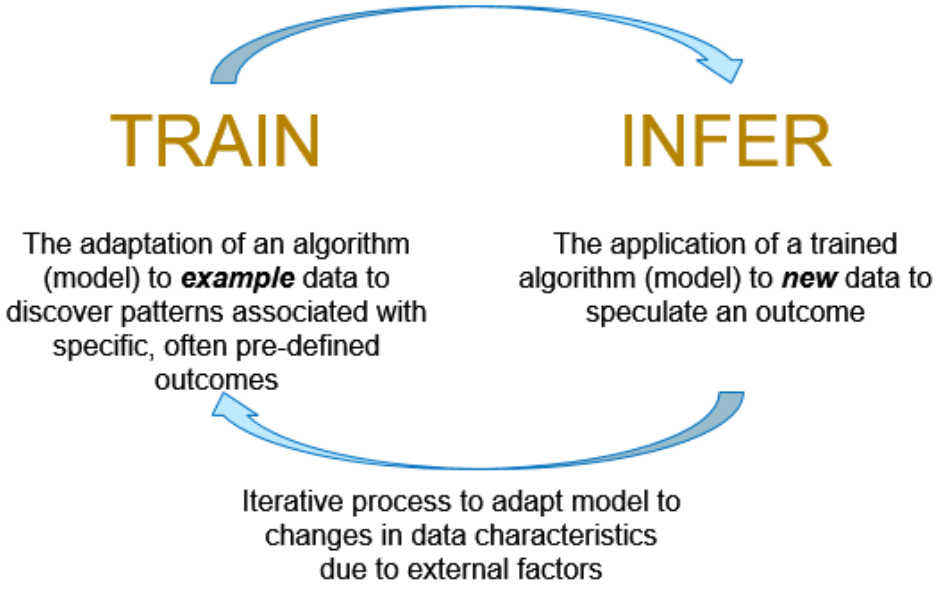


Figure 1. Deep Learning phases.

Deep Learning Inferencing

After a model is trained, the generated model may be *deployed* (forward propagation only) e.g., on FPGAs, CPUs or GPUs to perform a specific business-logic function or task such as identification, classification, recognition and segmentation [Figure 2].

The focus of this blog will be on the power of Dell EMC PowerEdge R7425 using NVIDIA T4-16GB GPUs to accelerate image classification and deliver high-performance inference throughput and low latency using various implementations of NVIDIA TensorRT™ an excellent tool to speed up inference and Malong RetailAI® software stack.

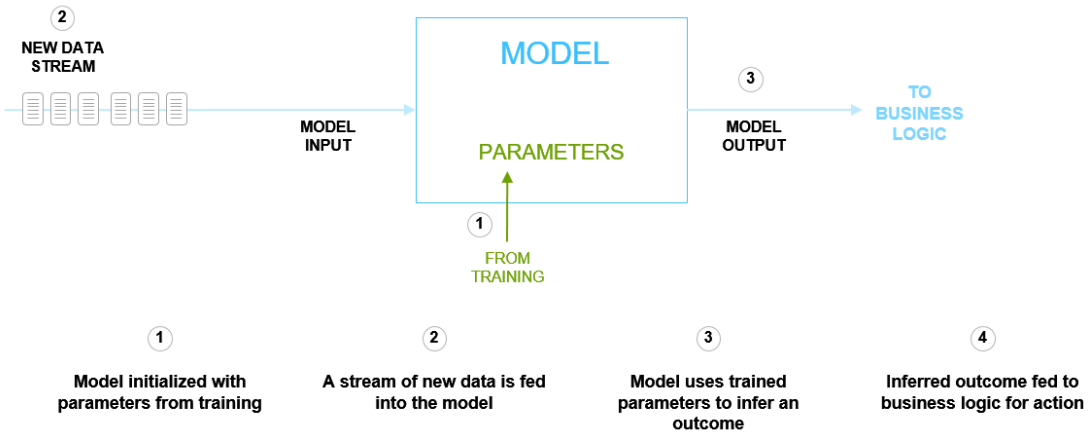


Figure 2. Inference Flow.

Why NVIDIA T4 GPU?

The NVIDIA® Tesla® T4 is single-slot, low profile, PCIE Express Gen3 Deep learning accelerator card based on the TU104 NVIDIA graphics processing unit (GPU). The NVIDIA T4 has 16GB GDDR6 memory and a 70W maximum power limit. It is a passively cooled board.

Tesla T4 is powered by NVIDIA Turing™ Tensor Cores to accelerate inference, video transcoding and virtual desktops. Turing Tensor Core technology with multi-precision computing for AI powers breakthrough performance from FP32 to FP16 to INT8, as well as INT4 precisions. It delivers up to 9.3X higher performance than CPUs on training and up to 36X on inference.

Figure 3: NVIDIA Tesla T4 GPU

Dell EMC PowerEdge R7425 Server

Dell EMC PowerEdge R7425-T4-16GB server supports the latest GPU accelerator to speed results in data analytics and AI applications, it enables fast workload performance on more cores for cutting edge applications such Artificial Intelligence (AI), High Performance Computing (HPC), and scale up software defined deployments. See Figure 4.

Figure 4. DELL EMC PowerEdge R7425

The Dell™ EMC PowerEdge™ R7425 is Dell's latest 2-socket, 2U rack server designed to run complex workloads using highly scalable memory, I/O, and network options The system features are based on AMD High performance processor, AMD SP3 which supports up to 32 AMD "Zen" x86 cores (AMD Naples Zeppelin SP3), up to 16 DIMMs, PCI Express® (PCIe) 3.0 enabled expansion slots, and a choice of OCP technologies.

The Dell EMC PowerEdge R7425 is a general-purpose platform capable of handling demanding workloads and applications, such as VDI cloud client computing, database/in-line analytics, scale up software defined environments, and high-performance computing (HPC).

The Dell EMC PowerEdge R7425 adds extraordinary storage capacity options, making it well-suited for data intensive applications that require greater storage, while not sacrificing I/O performance.

Why Malong?

Malong has built scalable product recognition algorithms based on weakly supervised deep learning. This technology has been applied successfully to identifying scan-fraud at self-checkout terminals ("SCOs") and staffed lanes, which are among the leading causes of shrinkage in brick-and-mortar retail environments.

Malong invented an algorithm called Curriculum Net [3], which is a breakthrough approach in learning directly from noisy and unbalanced visual data, which is exactly the kind of data seen in chaotic retail environments. Using this technique, the Malong RetailAI® system can learn without the need for additional human supervision in annotation, which is infeasible in large-scale retail scenarios.

Note that the Malong algorithm was put to the test in a worldwide image recognition competition held by Google Research called WebVision at CVPR, the premier conference in computer vision, with participation by over 100 scientific research organizations in the challenge. The Malong system won first place by a wide margin, outperforming second place by a relative error rate of nearly 50%.

In scan fraud, one common behavior is often referred to as "mis-scan", or scan avoidance, in which case a shopper intentionally or accidently fails to scan the barcode. The other is called "ticket-switching", in which case a customer scans a cheaper product's barcode while taking a more expensive item. These problems are difficult to detect using conventional means, even compared to direct human observation; leading to significant losses for retailers. Traditional approaches to solving this problem often lead to

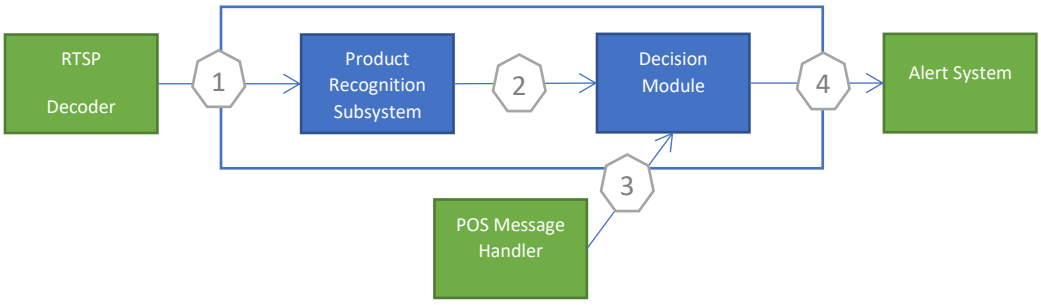
large percentages of false positives, which cause inconvenience to shoppers and add operational overhead to the business.

Malong addresses these problems by leveraging novel computer vision algorithms, to perform intelligent video analytics (IVA) for loss prevention at large scale, by accurately discovering mis-scans or ticket-switching at SCOs and staffed lanes in near real time. Scalability is key. As retailers may have many stores, huge store formats, and massive numbers of SKUs, it would be infeasible to require numerous deep learning models for each store and for different groups of products. The solution notably uses only one set of models, the same for all stores and for all products a retailer may have.

The solution runs on a Dell EMC PowerEdge R7425 containing NVIDIA T4 GPUs, which is physically located in a store. It will receive RTSP feeds from IP cameras installed above the registers at 30FPS. These feeds are channeled to the solution to ultimately infer if mis-scan or ticket-switching activity is occurring. A second feed required by the solution is access to the corresponding point-of-sale (POS) messages.

The Malong RetailAI® solution is built on NVIDIA Metropolis, which includes NVIDIA's rich software development kits, including NVIDIA DeepStream and NVIDIA TensorRT, and its underlying libraries. It is containerized using the NVIDIA Container Toolkit, which runs GPU accelerated Docker containers. The key components of the video streaming pipeline are an RTSP video decoder, POS message handler, alert system and weakly-supervised deep learning-based computer vision algorithms which implement scalable product recognition.

The number maps the input and the output:



1. The Malong video pipeline runs primarily on the GPU. From the moment a frame enters the system, it is decoded on the GPU and processing stays within the GPU for the remainder of the pipeline. The T4 provides super real-time, fully accelerated hardware-based video decoding. The frame buffer has been already pinned in GPU memory since it is created, which is consumed by GPU inference in-place later; therefore, there isn't an extra cost of moving the buffer between CPU and GPU.
2. The product recognition module contains a subsystem that identifies and tracks products spatial temporally, even as they are obscured temporarily, through re-identification. That is, the task is to accurately identity the same unique product across multiple frames and link the same

products into a trace. Every trace is represented as a sequence of binary elements. The 1st part in an element is a "location" within the field of view, the other part is a "timestamp" indicating approximately when this product moved to this location.

3. The POS message handler will interface with the register messaging system to incorporate into the processing pipeline. Each scan signal must contain when the scan happened and what product scanned. If a customer tries to scan a wine with a wax candle barcode, then a signal with candle barcode will be sent to the decision module.
4. The decision module will decide whether the current scan is a possible scan fraud event, e.g. ticket-switching or mis-scan by inferring based on the following pieces of information:

    i. The traces including inferred products, their movement and trajectory,

    ii. The physical context of the register in the field of view,

    iii. The register POS messages, including their scan timing and scan products.

    Note: the decision module only considers the physical objects in the scene; that is, it in no way whatsoever considers the people in the scene. This ensures customer privacy and no bias, a fundamental principle in responsible AI.

All models in the pipeline are heavily optimized with NVIDIA TensorRT to best leverage the NVIDIA T4 Turing architecture Tensor Cores with mixed precision accelerated inference, which significantly increases throughput and efficiency on the Dell EMC PowerEdge R7425. These optimizations provide for a 480%+ speed up when compared to not using TensorRT on the same hardware. Compared to running on CPU only, the difference with an optimized GPU version is a 99%+ reduction in processing time.

Processor	Speed of Core Model	% Difference to Optimized
CPU	4.7 seconds	-99.36%
NVIDIA T4 GPU (without optimizations)	0.175 seconds	-82.8%
NVIDIA T4 GPU with TensorRT optimizations	0.03 seconds	N/A

In terms of how many SCOs can be supported via the Malong RetailAI® solution:

1. Simultaneously, 4-6 SCOs can be supported per single T4, depending on specific configuration settings.
2. Non-simultaneously, effectively an unlimited number. This is relevant because all the SCOs in a store are typically not active all at once, all the time. So, in theory, all the SCOs can be mapped to a single T4, but when the concurrency reaches maximum, a random subset of SCOs will be covered.

Conclusions

The Dell EMC PowerEdge R7425 using AMD EPYC processor is a powerful platform in conjunction with NVIDIA T4 GPU running the Malong container stack, it provides end-to-end retail analytic capability.

Leveraging a Dell EMC PowerEdge server in a store with Malong RetailAI®, brings to bear industry-leading computer vision technology to help solve retail business problems. In this solution, retailers stand to benefit from significant reductions in shrink, minimized inventory data accuracy issues (ticket-switching, for example, will cause inventory numbers to not match what was sold), and increased revenue (through re-scans).

Appendix- PowerEdge R7425-T4-16GB Server – GPU Features

PowerEdge R7425-T4-16GB - GPU Features

Server		R7425-T4
CPU		
	CPU model	AMD EPYC 7551 32-Core Processor
GPU		
	GPU model	Tesla T4-16GB
	GPU Architecture	NVIDIA Turing
	Attached GPUs	6
Features per GPU		
	Driver Version	410.79
	Compute Capability	7.5
Multiprocessor		
	Multiprocessors (MP)	40
	CUDA Cores/MP	64
	CUDA Cores	2,560
	Clock Rate (GHz)	1.59
Memory		
	Global Memory Bandwidth (GB/s)	300

Server	R7425-T4
Global Memory Size (GB)	16
Constant Memory Size (KB)	65
L2 Cache Size (MB)	4
Bus Interface PCIe	
Generation	3
Link Width	16
Peak Performance Floating Point Operations (FLOP) and TOPS	
Single-Precision - FP32 (Teraflop/s)	8.1
Mixed Precision - FP16/FP32 (Teraflop/s)	65
Integer 8 - INT8 (Tera Operations /s)	130
Integer 4 – INT4-16GB (Tera Operations /s)	260
Power	
Min Power Limit (W)	60
Max Power Limit (W)	70

References

[1] H. Zhao, O. Gallo, I. Frosio and J. Kautz, "Loss Functions for Image Restoration with Neural Networks," in IEEE Transactions on Computational Imaging, vol. 3, no. 1, pp. 47-57, March 2017

[2] L. J. Buturovic and L. T. Citkusev, "Back propagation and forward propagation," [Proceedings 1992] IJCNN International Joint Conference on Neural Networks, Baltimore, MD, 1992, pp. 486-491 vol.4.

[3] S. Guo, W. Huang, H. Zhang, C. Zhuang, D. Dong, M. R. Scott, and D. Huang. “Curriculum Net: Weakly Supervised Learning from Large-Scale Web Images,” in the European Conference on Computer Vision (ECCV), September 2018.