SIX STEPS FOR MEANINGFUL DEEP LEARNING OUTCOMES

70% of CIOs will aggressively apply data and artifical intelligence (AI) by 2021[1]

Are you ready?

STEPS FOR DEPLOYING DEEP LEARNING

1 DATA

Identifying, characterizing, consolidating, and optimizing data lays the foundation for effective AI modeling and training.

2 TOOLS

Software tools and libraries (like TensorFlow* and Apache Spark*) can help jump start the creation, training and deployment of AI models.

- Containers
- or
- Intel-optimized frameworks
- or
- Libraries

3 TRAINING

Deep learning needs an initial model from which to build a neural network. With training, the model increases its ability to identify correlations among disparate data.

- Cloud
- or
- Multi-use cluster (Spark*/HPC/other)
- or
- Dedicated training cluster (with Nauta*)

4 MODELING AND OPTIMIZATION

Training a model for accuracy may require huge volumes of data. HPC can help accelerate this compute-intensive step. Once the neural network's accuracy has been verified, the trained model is ready for use.

- Trained model weights
- Intel® Distribution of OpenVINO™ toolkit

5 QUANTIZATION

Quantization involves intelligent data compression for use with a neural network. Reducing the size of the working data sets reduces compute demands and accelerates a model's throughput for faster results.

6 INFERENCE

Inference is the process of using a trained model to return meaningful information from new or existing data sets. An AI model can reside on-premise or hosted in a cloud or edge-based solution.

- Cloud
- or
- Multi-use cluster (Spark*/HPC/other)
- or
- Edge server
- or
- Device

OPTIMIZE YOUR HPC ENVIRONMENT FOR AI WITH INTEL

FASTER TIME TO INSIGHT[2]

UP TO 30x

Start introducing AI capabilities today using your existing Intel® architecture. For example, the 2nd generation Intel® Xeon® Scalable processor with Intel® Deep Learning Boost offers up to 30x faster time to insight[2] vs. previous generation Intel Xeon processors.

Want to learn more? Read the white paper *Jump-Start Your AI Journey With Your Existing HPC infrastructure*

Read the white paper

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.

Performance results are based on testing as of the date set forth in the configurations and may not reflect all publicly available security updates. See configuration disclosure for details. No product or component can be absolutely secure.

[1] IDC: 70 percent of CIOs will aggressively apply data and AI to IT operations, tools, and processes by 2021 IDC Reveals Worldwide CIO Agenda 2019 Predictions

[2] 30x inference throughput improvement on Intel® Xeon® Platinum 9282 processor with Intel® DL Boost: Tested by Intel as of 2/26/2019. Platform: Dragon rock 2 socket Intel® Xeon® Platinum 9282 (56 cores per socket), HT ON, turbo ON, Total Memory 768 Gb (24 slots/ 32 Gb/ 2933 MHz), BIOS:SE5C620.86B.0D.01.0241.112020180249, CentOS 7 Kernel 3.10.0-957.5.1.el7.x86_64, Deep Learning Framework: Intel® Optimization for Caffe* version: https://github.com/intel/caffe d554cbf1, ICC 2019.2.187, MKL DNN version: v0.17 (commit hash: 830a10059a018cd2634d94195140cf2d8790a75a), model: https://github.com/intel/caffe/blob/master/models/intel_optimized_models/int8/resnet50_int8_full_conv.prototxt, BS=64, No datalayer syntheticData:3x224x224, 56 instance/2 socket, Datatype: INT8 vs. Tested by Intel as of July 11, 2017: 2S Intel® Xeon® Platinum 8180 CPU @ 2.50 GHz (28 cores), HT disabled, turbo disabled, scaling governor set to “performance” via intel_pstate driver, 384 Gb DDR4-2666 ECC RAM. CentOS Linux* release 7.3.1611 (Core), Linux kernel 3.10.0-514.10.2.el7.x86_64. SSD: Intel® SSD DC S3700 Series (800 Gb, 2.5in SATA 6 Gb/s, 25nm, MLC). Performance measured with: Environment variables: KMP_AFFINITY='granularity=fine, compact‘, OMP_NUM_THREADS=56, CPU Freq set with cpupower frequency-set -d 2.5G -u 3.8G -g performance. Caffe: (http://github.com/intel/caffe/), revision f96b759f71b2281835f690af267158b82b150b5c. Inference measured with “caffe time --forward_only” command, training measured with “caffe time” command. For “ConvNet” topologies, synthetic data set was used. For other topologies, data was stored on local storage and cached in memory before training. Topology specs from https://github.com/intel/caffe/tree/master/models/intel_optimized_models (ResNet-50). Intel C++ compiler ver. 17.0.2 20170213, Intel® MKL small libraries version 2018.0.20170425. Caffe run with “numactl -l“.

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.

0719/KR/CAT/PDF

340928-001EN