WHITE PAPER

Artificial Intelligence
Intel® AI Builders

Teaching Machines to do Image Classification in Health and Life Sciences: Intel® Xeon® Scalable Processors in Lab Coats

Authors

Kyle Ambert
Sr. Deep Learning Data Scientist, Artificial Intelligence Products Group

Deepthi Karkada
Data Scientist, Artificial Intelligence Products Group

Krishna Sumanth Muppalla

Kushal Datta
Research Scientist, Intel AI Lab, Artificial Intelligence Products Group

Introduction

At Intel, we're quite interested in how systems can be made smarter to solve meaningful tasks relevant to healthcare providers and patients today. The Intel and MobileODT Cervical Cancer Screening Kaggle competition, for example, challenged data scientists to train our respective computational systems to assist with the identification of early-stage cervical cancer in medical images. Inspired by this project, let's take a closer look at deep learning systems for image recognition in health and life sciences (HLS) and how Intel's portfolio of products for artificial intelligence (AI) is helping make HLS AI solutions a reality.

Unpacking Deep Learning

Production deep learning systems have two main categories of computation that researchers like to discuss: training and inference. Training is the process by which our system finds patterns in data. Inference is the process of using the trained model to make predictions about data we have not previously seen. To better understand training for deep learning systems, it is helpful to contrast it with the same process for traditional supervised machine learning.

In traditional supervised machine learning, systems require an expert to use his or her domain knowledge to specify the information (called features) in the input data that will best lead to a well-trained system. For example, in the case of an image classification problem wherein one wishes to classify health records associated with patients with a history of heart disease versus everyone else, they might direct the system to look for a history of Coumadin* prescriptions or insurance billing codes associated with cardiac events. The system would use this information, along with the expert-assigned labels indicating which patient records are truly associated with incidence of heart disease, to build a statistical model that maps the input features to the labels.

Deep learning systems are slightly different. Rather than specifying the features in our data that we think will lead to the best classification accuracy, we let the machine find this information on its own. Often, it is able to look at the problem in a way that even an expert wouldn't have been able to imagine. To do this, it is the job of the data scientist to create a neural network architecture that is best able to find these features and patterns in the data. During training, we pass data through the neural network, error-correct after each sample, and iterate until the best network parametrization is achieved. After the network has been trained, the resulting architecture can be used for inference.

Neural network algorithms created via deep learning training are often more versatile than machine learning systems that have been manually crafted by experts. Nowhere is this contrast more evident than in the case of image classification. Although humans are generally quite capable at the typical image

classification types of problems, they're not always very good at articulating the rationale behind their decisions in a way that is useful to a classification algorithm. For example, every day radiologists make clinical judgements about whether a patient study is symptomatic of cancer, but it would be difficult for them to articulate a general heuristic for identifying all cancers in any images associated with any patient, agnostic of image quality or image rotation. In a sense, this is what a deep learning system is able to learn, given sufficient data.

Deep Learning Training on Large Images

If you were to spend much time in the deep learning image classification literature, you'd probably start to notice a pattern in the data sets that are typically discussed. MNIST, ImageNet, CIFAR-10—these are the data sets frequently used for image classification benchmarking, and they all contain images that are quite small compared to the sorts of images acquired in health and life sciences, especially in the case of high-content microscopy, which captures 16-bit images. Such image capture methods translate to data on the order of 10-12MB. This may not seem that large relative to the types of big data problems researchers like to talk about these days, but, considering that deep learning systems are iteratively trained over batches of data that could range from groups of eight images to groups of 1200 images, it adds up.

Due to their support for greater memory footprints, CPU-based deep learning systems are uniquely equipped to handle the memory demand associated with training a neural network on large images and accommodating the size of the image batches. In research my team recently presented at SC17, we demonstrated that a CPU-based system could handle a memory footprint in excess of 40GB for a real-world microscopy classification task (Figure 1).

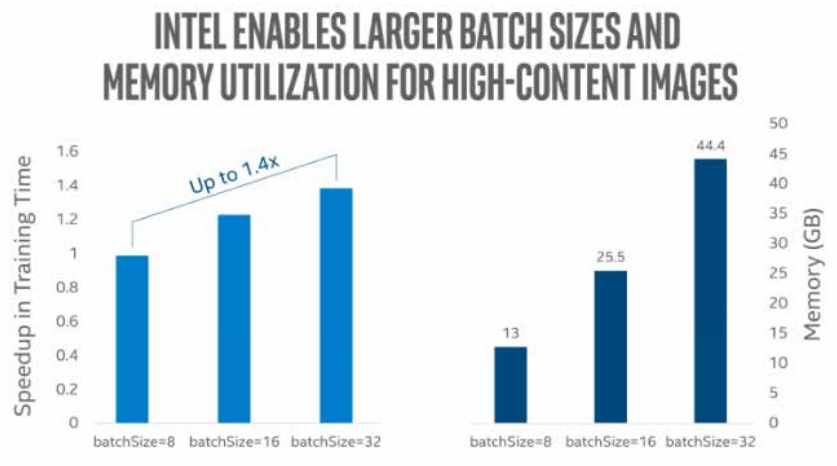

Figure 1: Speedup in training time increases with batch size as does size of MCNN workload supported on single node Intel® Xeon® Gold 6148 processor. Increasing image batch size shortens training time (in exchange for decreased accuracy and convergence). However, larger batch sizes also increase memory requirements. When working with high-content images, even moderately-sized workloads can lead to substantial increases in the memory footprint associated with learning a network used in the real world.[2]

If we need to scale training further, we could additionally look to Intel® Xeon Phi™ processors, which NERSC and Stanford chose for their 15 petaflop system.

Deep Learning Data Considerations in HLS

If you are wondering, "Where will the images needed to train our algorithm come from?", it's actually one of the most complicated and time-consuming parts of deep learning today. To train a deep learning neural network, most commonly-used methods today will require labeled images. "Labeled," in this context, means that an expert has reviewed and assigned an annotation (e.g., cancer vs. not cancer) to the images in our training data. To this end, there may be publicly available databases of labeled images that could be used to train a system like this (e.g., Medical Image Net, from Stanford Medicine Radiology Informatics' Langlotzlab), or, if you work at a healthcare institution, there are likely to be internal resources you'll be able to use. Making use of these resources, or acquiring new data at scale, does require a carefully planned and executed data strategy (for more guidance, please see my earlier post, "Key Considerations for Building a Robust Data Strategy").

Deep Learning Inference: Speed and Efficiency

Once trained, our deep learning system must rapidly and efficiently deliver inference results to the clinician it's supporting. Intel's AI solution portfolio offers many great options for inference. Intel Xeon Scalable processors make a great foundation for inference in our theoretical HLS image classification system, as they enable inference to be run atop the same Intel® architecture already relied upon for workloads like advanced analytics. To cite an example from another field, Alibaba uses Intel Xeon Scalable processors for inference in their Alibaba Store Concierge intelligent customer service chatbot, realizing

up to 80% gen-on-gen performance improvements in some business environments[3]. Intel FPGA accelerators further extend inference power efficiency in the datacenter compared with Intel Xeon processors alone[4]. Due to this, Microsoft employs a combination of Intel Xeon processors and Intel FPGAs to support AI workloads in its Azure Cloud platform[5]. FPGAs could similarly increase power efficiency in our example HLS image classification system.

More Potential for Deep Learning

Using AI to process medical images has the potential to improve treatment and positively impact countless lives, but the challenges discussed here are only a small subset of a larger group of concerns that must be addressed for an AI solution to be made real. Regardless, AI's potential makes me excited to continue Intel's work to solve these problems.

For more information on deep learning networks for image recognition, check out "Convolutional Neural Networks, Part 1: Historical Significance."

[1]Benchmark results were obtained prior to implementation of recent software patches and firmware updates intended to address exploits referred to as "Spectre" and "Meltdown". Implementation of these updates may make these results inapplicable to your device or system.

Workload: Image set BBBC021 : Human MCF7 Cells – compound profiling experiment. Godinez et al, A multi-scale convolutional neural network for phenotyping high-content cellular images. *Bioinformatics*, 2017.

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 names and brands may be claimed as the property of others

Intel Cluster Configuration Details

1x **R2208WTTYS Intel® Xeon® Processor Server System** *(Master Node)*

- 2x Intel® Xeon® Processor E5-2699 v4
- 256GB DDR4 RAM
- 1x Intel® Omni-Path Fabric x16 PCIe card

4x **HNS7200AP** Chassis

- 16x **Intel® Xeon Phi™ Processor 7290F** (total)
- 192GB DDR4 RAM (each)
- 1x 1.6TB Intel® SSD DC S3610 Series SC2BX016T4 (each)
- 1x 480GB Intel® SSD DC S3520 Series SC2BB480G7 (each)
- 1x Intel® Omni-Path Fabric x16 PCIe card (each)

2x S2P2SY1Q Intel® Xeon® Processor based Server Systems

- 2x **Intel® Xeon® Scalable Gold 6148 processors** (each)
- 192GB DDR4 RAM (each)
- 1x 1.6TB Intel® SSD DC S3610 Series SC2BX016T4 (each)
- 1x 480GB Intel® SSD DC S3520 Series SC2BB480G7 (each)
- 1x Intel® Omni-Path Fabric x16 PCIe card (each)

1x 42U server rack cabinet

1x Intel® Omni-Path Director Class Switch (48 port 100GB)

1x Extreme Networks* 48 port 10GB switch

1x Catalyst* 24 port 1GB switch

Operating System: CentOS Linux release 7.2

Acceleration Library: Intel® MKL 2017/DAAL/Intel Caffe*, Python* 2.7.5 (default, Aug 4 2017, 00:39:18), Tensorflow* 1.3.0.rc0 (Aug 3rd, 2017)

[2] Workload: Image set BBBC021 : Human MCF7 Cells – compound profiling experiment. Godinez et al, A multi-scale convolutional neural network for phenotyping high-content cellular images. Bioinformatics, 2017.

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 names and brands may be claimed as the property of others

Intel Cluster Configuration Details.

1x **R2208WTTYS Intel® Xeon® Processor Server System** *(Master Node)*

2x Intel® Xeon® Processor E5-2699 v4

256GB DDR4 RAM

1x Intel® Omni-Path Fabric x16 PCIe card

4x **HNS7200AP** Chassis

- 16x **Intel® Xeon Phi™ Processor 7290F** (total)
- 192GB DDR4 RAM (each)
- 1x 1.6TB Intel® SSD DC S3610 Series SC2BX016T4 (each)
- 1x 480GB Intel® SSD DC S3520 Series SC2BB480G7 (each)
- 1x Intel® Omni-Path Fabric x16 PCIe card (each)

2x S2P2SY1Q Intel® Xeon® Processor based Server Systems

- 2x **Intel® Xeon® Scalable Gold 6148 processors** (each)
- 192GB DDR4 RAM (each)
- 1x 1.6TB Intel® SSD DC S3610 Series SC2BX016T4 (each)
- 1x 480GB Intel® SSD DC S3520 Series SC2BB480G7 (each)
- 1x Intel® Omni-Path Fabric x16 PCIe card (each)

1x 42U server rack cabinet

1x Intel® **Omni-Path Director Class Switch** (48 port 100GB)

1x Extreme Networks* 48 port 10GB switch

1x Catalyst* 24 port 1GB switch

Operating System: CentOS Linux release 7.2

Acceleration Library: Intel® MKL 2017/DAAL/Intel Caffe*, Python* 2.7.5 (default, Aug 4 2017, 00:39:18), Tensorflow* 1.3.0.rc0 (Aug 3rd, 2017)

Acknowledgements: Novartis Institutes for BioMedical Research (Basel & Cambridge): Michael Derby, William J. Godinez, Imtiaz Hossain, Stephen Litster, Michael Steeves, Xian Zhang, Wolfgang Zipfel. Intel Corporation: Kyle Ambert, Joe Bailey, Clayton Craft, Kushal Datta, Michael Demshki, Sheng Fu, Deepthi Karkada, Kristina Kermanshahche, Adam Procter, Jon Markee, Patrick Messmer, Krishna S. Muppalla, Rakib Sarwar, Vikram Saletore, Oleg Stepanov, Nikita Swinnen-Galbraith.

[3] https://itpeernetwork.intel.com/xeon-scalable-powers-future-ai/

[4] https://www.altera.com/solutions/technology/artificial-intelligence/overview.html

[5] https://itpeernetwork.intel.com/xeon-scalable-powers-future-ai/