Intel's optimized tools and frameworks for machine learning and deep learning

This article gives an introduction to the Intel's optimized machine learning and deep learning tools and frameworks and also gives a description of the Intel's libraries that have been integrated into them so they can take full advantage and run fastest on Intel® architecture. This information will be useful to first-time users, data scientists, and machine learning practitioners, for getting started with Intel optimized tools and frameworks.

Introduction

Machine learning (ML) is a subset of the more general field of artificial intelligence (AI). ML is based on a set of algorithms that learn from data. Deep learning (DL) is a specialized ML technique that is based on a set of algorithms that attempt to model high-level abstractions in data by using a graph with multiple processing layers (https://en.wikipedia.org/wiki/Deep_learning).

ML, and in particular DL, are currently used in a growing number of applications and industries, including image and video recognition/classification, face detection, natural language processing, and financial forecasting and prediction.

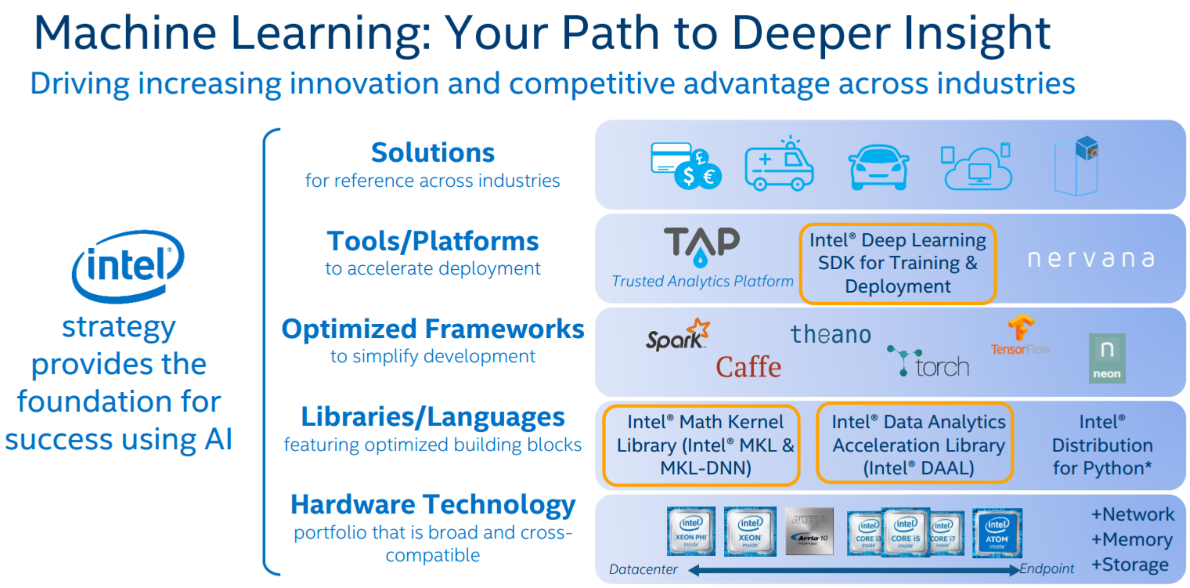
A convenient way to work with DL is to use the Intel's optimized ML and DL frameworks. Using Intel optimized tools and frameworks to train and deploy deep networks guarantees that these tools will use Intel® architecture in the most efficient way. Examples of how some of these frameworks have been optimized to take advantage of Intel architecture, as well as charts showing speed-up of these optimized frameworks compared to non-optimized ones, can be found in https://software.intel.com/en-us/ai/deep-learning

(See for example, http://itpeernetwork.intel.com/myth-busted-general-purpose-cpus-cant-tackle-deep-neural-network-training/

and

http://itpeernetwork.intel.com/myth-busted-general-purpose-cpus-cant-tackle-deep-neural-network-training-part-2/).

In the figure, we can see that the Intel's solution stack for ML and DL spans different layers. On top of the hardware, Intel has developed highly optimized math libraries that make the most efficient use of the several families of Intel® processors. Those optimized math libraries are the foundation for higher-level tools and frameworks that are used to solve ML and DL problems across different domains.



Machine Learning: Your Path to Deeper Insight

Driving increasing innovation and competitive advantage across industries

intel strategy provides the foundation for success using AI

Solutions
for reference across industries

Tools/Platforms
to accelerate deployment

TAP
Trusted Analytics Platform

Intel® Deep Learning SDK for Training & Deployment

nervana

Optimized Frameworks
to simplify development

Spark
Caffe
theano
torch
TensorFlow
neon

Libraries/Languages
featuring optimized building blocks

Intel® Math Kernel Library (Intel® MKL & MKL-DNN)

Intel® Data Analytics Acceleration Library (Intel® DAAL)

Intel® Distribution for Python*

Hardware Technology
portfolio that is broad and cross-compatible

Datacenter ↔ Endpoint

+Network
+Memory
+Storage

In the next section, a brief summary of Intel's libraries and tools used to optimize the frameworks will be described. Although these libraries and tools have been used to optimize the ML and DL frameworks for Intel architecture, they can also be used in other applications or software packages that require highly optimized numerical routines that can take advantage of the vectorization, multithreading, and distributed computing capabilities present in Intel® hardware.

Intel® software tools for machine learning and deep learning

Intel is actively working with the open source community to ensure that existing and new frameworks are optimized to take advantage of Intel architecture and is optimizing these ML and DL tools by using powerful libraries that provide building blocks to accelerate these tasks.

Intel has developed three libraries that are highly optimized to run on Intel architecture.

- Intel® Math Kernel Library (Intel® MKL) (https://software.intel.com/en-us/intel-mkl) includes a set of highly-optimized performance primitives for DL applications (https://software.intel.com/en-us/node/684759). This library also includes functions that have been highly optimized (vectorized and threaded) to maximize performance on each Intel® processor family. These functions have been optimized for single-core vectorization and cache memory utilization, as well as with automatic parallelism for multi-core and many-core processors.

  Intel MKL provides standard C and Fortran APIs for popular math libraries like BLAS, LAPACK, and FFTW, which means no code changes are necessary. Just relinking the application to use Intel MKL will maximize performance on each Intel processor family. This will provide great performance in DL applications with minimum effort.

Intel MKL is optimized for the most recent Intel processors, including Intel® Xeon® and Intel® Xeon Phi™ processors. In particular, it is optimized for Intel® Advanced Vector Extensions 2 and Intel® Advanced Vector Extensions 512 ISAs.

Intel® MKL library can be downloaded for free via the Community Libraries program (https://software.intel.com/sites/campaigns/nest/).

- Intel® Math Kernel Library for Deep Neural Networks (Intel® MKL-DNN) (https://01.org/mkl-dnn) is an open source performance library for DL applications that can be used to maximize performance on Intel architecture.

  This library provides optimized deep neural network primitives for rapid integration into DL frameworks. It welcomes community contributions for new functionality, which can be immediately used in applications ahead of Intel MKL releases. This way, DL scientists and software developers can both contribute to and benefit from this open source library.

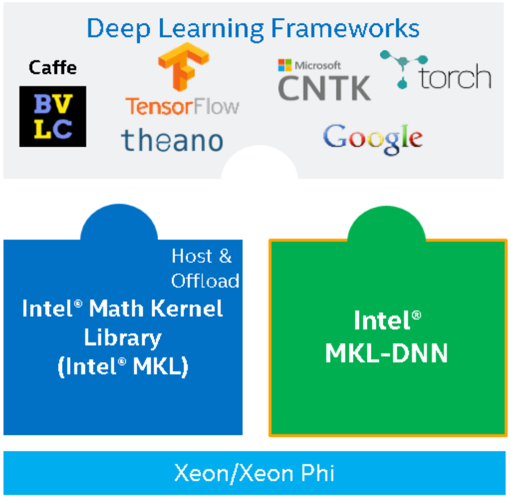
- Intel® Data Analytics Acceleration Library is a performance library of highly optimized algorithmic building blocks for all data analysis stages (preprocessing, transformation, analysis, modeling, validation, and decision making). It is designed for use with popular data platforms including Hadoop*, Spark*, R*, and others, for efficient data access. This library is available for free via the Community Licensing for Intel® Performance Libraries (https://software.intel.com/sites/campaigns/nest/).

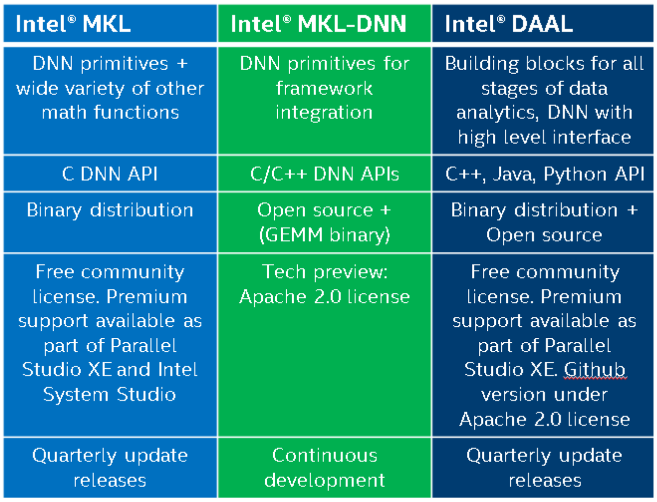
Deep learning frameworks

Intel's optimized ML and DL frameworks use the functionality of the libraries described in the previous section. They allow us to perform training and inference in a highly efficient manner using Intel processors.

Intel is actively working in integrating the math libraries into the various frameworks so the users of these frameworks can run their DL training and inference tasks in the most efficient way on Intel processors. For example, Intel® Distribution of Caffe* and Intel® Optimization for Theano* are integrated with the most recent versions of Intel MKL. Intel is also adding multimode capabilities to those frameworks in order to distribute the training workload across nodes, which results in reducing the total training time.

The interaction between the different libraries and frameworks previously described can be represented visually in the following block diagram, which shows how Intel MKL and Intel MKL-DNN libraries are used as building blocks for the several optimized frameworks.





Intel® MKL	Intel® MKL-DNN	Intel® DAAL
DNN primitives + wide variety of other math functions	DNN primitives for framework integration	Building blocks for all stages of data analytics, DNN with high level interface
C DNN API	C/C++ DNN APIs	C++, Java, Python API
Binary distribution	Open source + (GEMM binary)	Binary distribution + Open source
Free community license. Premium support available as part of Parallel Studio XE and Intel System Studio	Tech preview: Apache 2.0 license	Free community license. Premium support available as part of Parallel Studio XE. Github version under Apache 2.0 license
Quarterly update releases	Continuous development	Quarterly update releases

When working with DL techniques, there are two main steps:

- Training: In this step we attempt to create a model based on labeled data (for example, labeled images)

- Inference (also known as scoring): Once a model has been created, this model can be deployed to make predictions on the data (for example, to find objects in unlabeled images)

The Intel optimized ML and DL frameworks allow maximum performance on both steps when running on Intel architecture.

In DL, it is important to use frameworks and tools that have been optimized for the underlying hardware, because DL tasks (either training or inference) require a large amount of computation to complete. Although many of the popular DL frameworks (like Caffe, Theano, and so on) are open source software, they are not optimized to run efficiently on Intel architecture. You will only get the high performance on Intel architecture when you use the versions of these frameworks that Intel has optimized.

Whether you are interested in learning about ML and DL or if you are a data scientist with specific ML or DL tasks to perform, you can benefit from using Intel optimized frameworks. The best way to start your exploration of Intel optimized tools is to visit the Intel® Developer Zone portal for AI, https://software.intel.com/ai, where you can get general information about Intel® technologies supported for AI.

To download Intel optimized frameworks, as well as install documentation and training, you can go to https://software.intel.com/ai/deep-learning.

This webpage contains links to GitHub* pages to download the optimized frameworks, as well as links to varied documentation, videos, and examples.

If you are a software developer and interested in creating or optimizing your own DL tools or frameworks, you can take a look at examples of how Intel's modern code expert engineers have optimized popular frameworks. One example is in https://software.intel.com/videos/getting-the-most-out-of-ia-using-the-caffe-deep-learning-framework.

There you can learn how modern code techniques have been used to optimize the popular Caffe* DL framework, and you can apply those techniques to analyze and optimize your own ML or DL tool or framework to run with maximum efficiency on Intel architecture.

In addition to the Intel-optimized frameworks, Intel is also releasing a DL SDK, which is an integrated environment that allows the user to visualize different aspects of the DL process in real time, as well as handle visual representations of the DL models. Intel plans to continue working on this high-level tool for DL. You can visit https://software.intel.com/en-us/deep-learning-sdk to get more information about this new tool.

Conclusion

The use of data analytics (in particular ML and DL) has become a competitive advantage in many industries. Given the fast pace at which new ML and DL tools are currently developed, it is important for ML practitioners and data scientists to take advantage of tools and frameworks that have been already optimized and tuned to the underlying hardware, instead of investing time and resources trying to optimize them or loosing advantage because of long processing times when using non-optimized tools. Intel is actively working with the open source community to optimize ML and DL tools and frameworks, which will offer maximum performance on Intel architecture with minimum effort from the user. To start using Intel's optimized tools for your ML and DL needs, visit https://software.intel.com/en-us/ai.

Notices

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. Check with your system manufacturer or retailer or learn more at intel.com.

No license (express or implied, by estoppel or otherwise) to any intellectual property rights is granted by this document.

Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade.

This document contains information on products, services and/or processes in development. All information provided here is subject to change without notice. Contact your Intel representative to obtain the latest forecast, schedule, specifications and roadmaps.

The products and services described may contain defects or errors known as errata which may cause deviations from published specifications. Current characterized errata are available on request.