ACCELERATING DEEP LEARNING WORKLOADS WITH INTEL® AI ANALYTICS TOOLKIT & 3RD GEN INTEL® XEON® SCALABLE PROCESSORS

Intel® AI Analytics Toolkit, Powered by oneAPI

Louie Tsai

Agenda

- Introduce Deep Learning components from Intel® AI Analytics Toolkit (AI Kit)
  - Intel® AI Analytics Toolkit
  - Intel-optimized DL frameworks overview
  - Model Zoo for Intel Architecture overview
  - **Tutorial** : Performance benefit of Intel Optimization for TensorFlow
    - Profile resnet50v1.5 model from Model Zoo on DevCloud among Stock and Intel-optimized TF
- Accelerating AI performance on 3rd Gen Intel® Xeon® with Intel-optimized TF and Bfloat16
  - Bfloat16 acceleration on 3rd Gen Intel® Xeon
  - **Tutorial** : performance benefit by using Bfloat16 acceleration on 3rd Gen Intel Xeon
    - Profile resnet50v1.5 model among different data type

Intel® AI Analytics Toolkit(beta)

Accelerate end-to-end AI and Data Analytics pipelines with frameworks and python tools optimized for Intel® architectures using oneAPI libraries

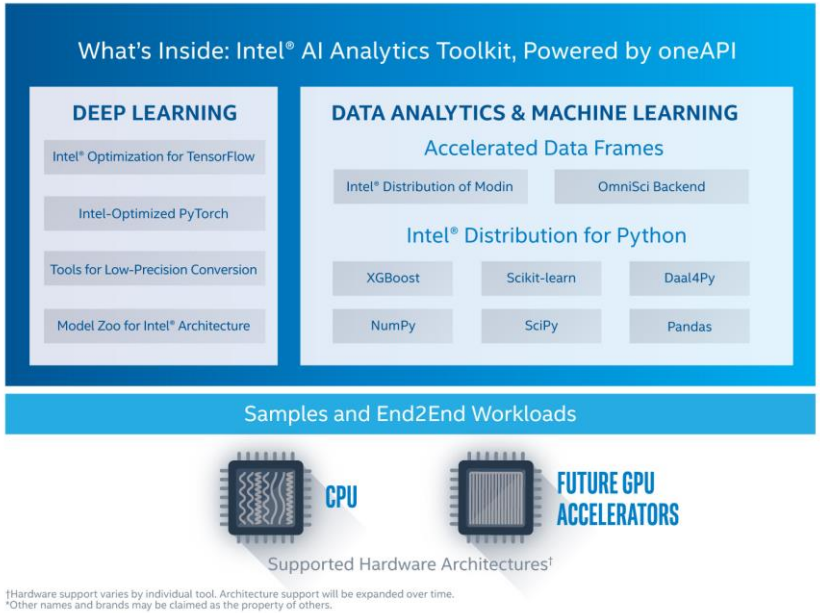
Who Uses It?

Data scientists, AI Researchers, ML and DL developers, AI application developers

Top Features/Benefits

Deep learning performance for training and inference with Intel optimized DL frameworks, tools for low precision and Model Zoo

Drop-in acceleration for Data analytics and Machine learning workflows with compute-intensive Python packages for easy scale out and performance



Watch introductory webinar on the AI Kit

[AI Analytics PART 1: Optimize End-to-End Data Science and Machine Learning Acceleration]

Benefits of Intel® AI Analytics Toolkit

Maximize the Power of Intel® XPU's for your AI and Analytics Pipelines

ACCELERATE PERFORMANCE

Get the latest Analytics and AI optimizations from Intel in one place; maximize performance on CPU's and future GPU/Xᵉ architectures via oneAPI libraries powering under the hood.

STREAMLINE END2END WORKFLOWS

Optimize and scale end-to-end workflows without the hassle of dependencies and need for external packages

IMPROVE PRODUCTIVITY

Alleviate the uncertainty associated with the Conda package manager through a version-controlled binary installation

SPEED DEVELOPMENT

Reduce the learning curve with drop-in replacement for Python packages with minimal to no code changes.

Get started quickly with samples, pre-trained models, and end-to-end workloads

DEEP LEARNING COMPONENTS OF INTEL® AI ANALYTICS TOOLKIT

Key DL Components

Intel® Optimization for TensorFlow

- In collaboration with Google, TensorFlow has been directly optimized for Intel® architecture (IA) using the primitives of Intel® oneAPI Deep Neural Network Library (oneDNN). The latest TF binary version compiled by setting CPU-related config (--config=mkl) is included as part of the toolkit.

PyTorch Optimized for Intel® Technology

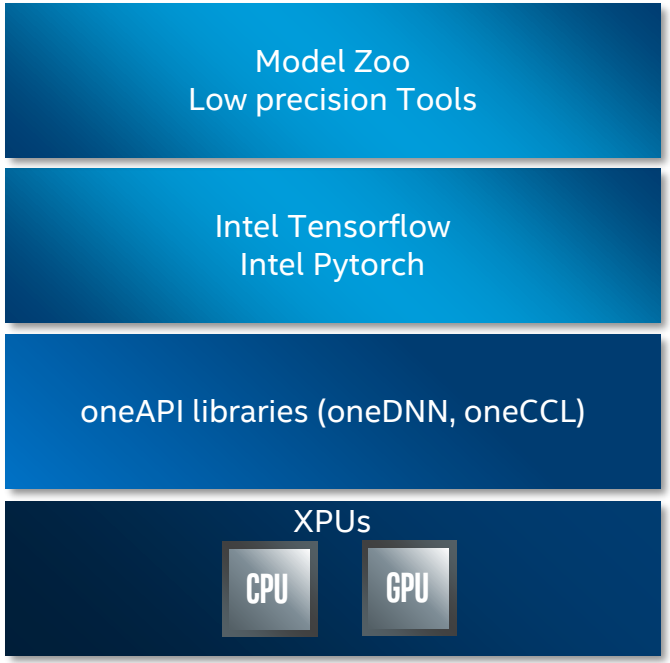
- In collaboration with Facebook, this popular framework is now directly combined with many Intel optimizations to provide superior performance on IA. Binary version of latest Pytorch release is included as part of the toolkit.

Tools for Low Precision Conversion

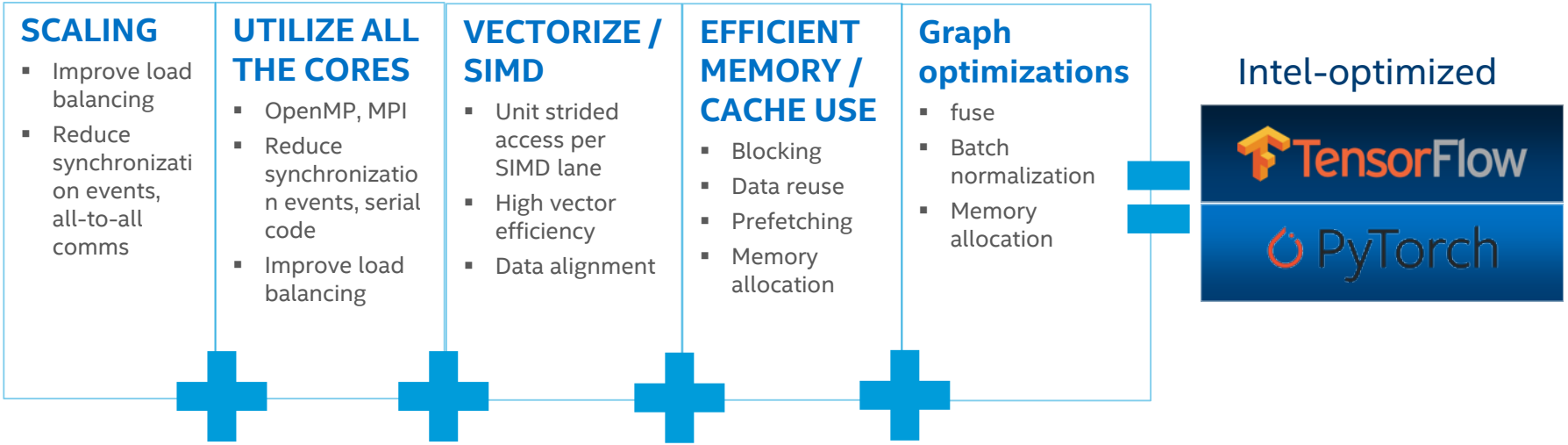
- The tool provides the capability to quantize models trained with FP32 precision to int8 precision and accelerate deep learning inference workflows.

Model Zoo for Intel® Architecture

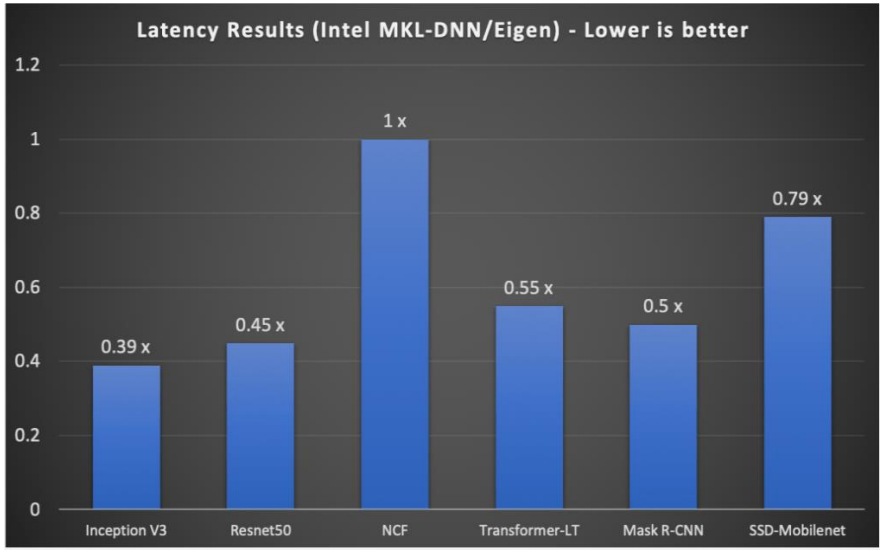
- Many popular open source machine learning models are pre-optimized by Intel to run efficiently on IA.

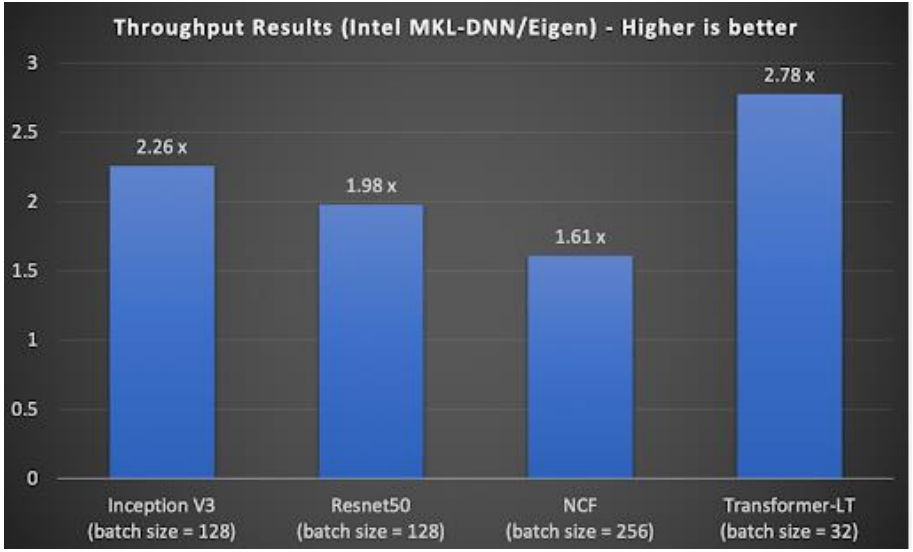


Deep Learning Frameworks, Optimized by Intel



Performance gains seen from Intel Tensorflow





We measured the throughput of ResNet-50 on a 2nd gen Intel Xeon Scalable processor (formerly codenamed Cascade Lake), more specifically Intel® Xeon® Platinum 9282 processor, a high core-count multi-chip packaged server multiprocessor, using Intel® Optimization for Caffe*. We achieved 7878 images per second by simultaneously running 28 software instances each one across four cores with batch size 11. The performance on NVIDIA Tesla V100 is 7844 images per second and NVIDIA Tesla T4 is 4944 images per second per NVIDIA's published numbers as of the date of this publication (May 13, 2019).

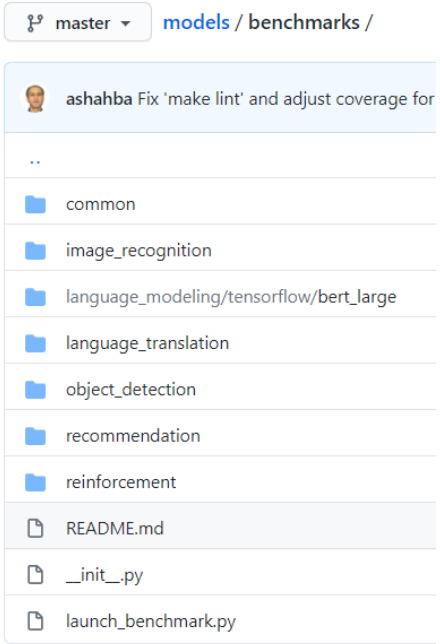
https://www.intel.com/content/www/us/en/artificial-intelligence/posts/improving-tensorflow-inference-performance-on-intel-xeon-processors.html

We demonstrated the effectiveness of Intel Xeon processors with optimized deep learning software, and achieved the throughput of ResNet-50 7878 images per second on Intel Xeon Platinum 9282 processors, outperforming NVIDIA's best GPUs.

https://software.intel.com/en-us/articles/intel-cpu-outperforms-nvidia-gpu-on-resnet-50-deep-learning-inference

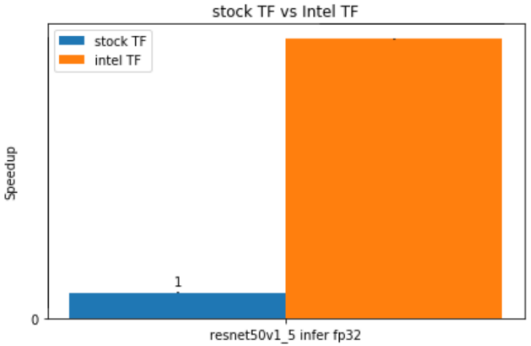
Model Zoo for Intel® Architecture (IA)

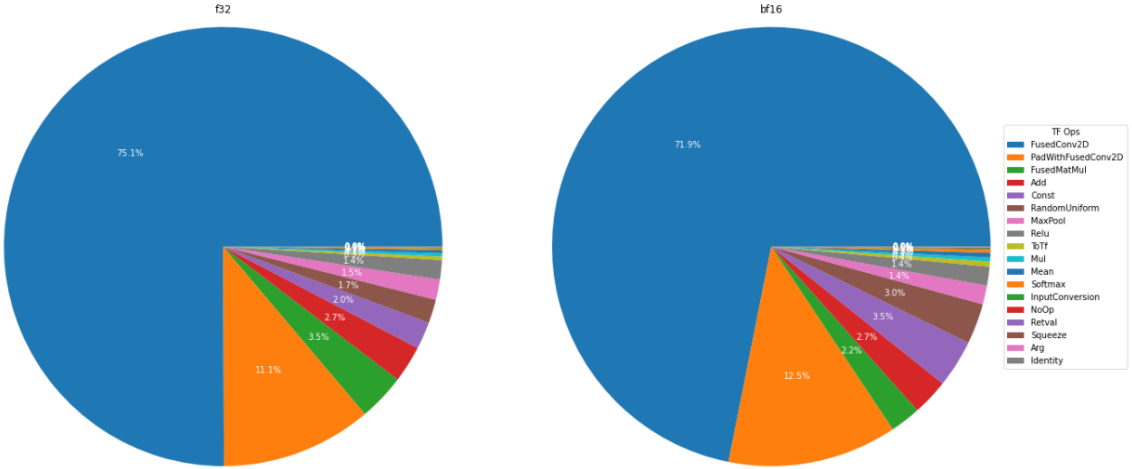
- **For many popular open-source machine learning models, Model Zoo has**
  - links to pre-trained models
  - sample scripts
  - best practices
  - step-by-step tutorials
- **Purpose of the Model Zoo**
  - Learn which AI topologies and workloads Intel has optimized to run on its hardware
  - Benchmark the performance of optimized models on Intel hardware
  - Get started efficiently running optimized models in containers or on bare metal



Demo : Performance benefit of Intel TensorFlow

- Profile resnet50v1.5 model from Model Zoo on the Intel® DevCloud and compare Stock TF vs Intel-optimized TF
  - Identify the **speedup** among different **TF operations** by using oneDNN library



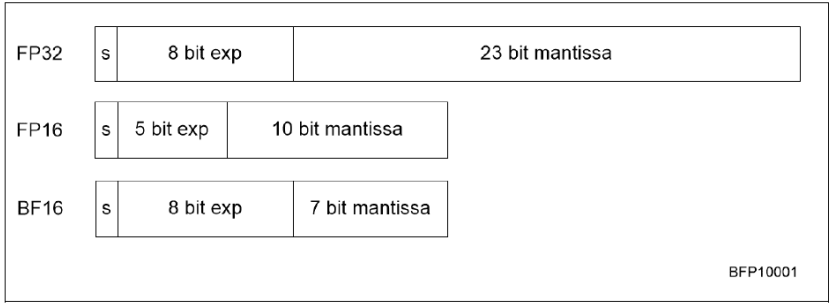


ACCELERATE AI PERFORMANCE WITH BFLOAT16 ON 3RD GEN INTEL® XEON® PROCESSORS

Bfloat16 Floating-point Format

BF16 has several advantages over FP16:

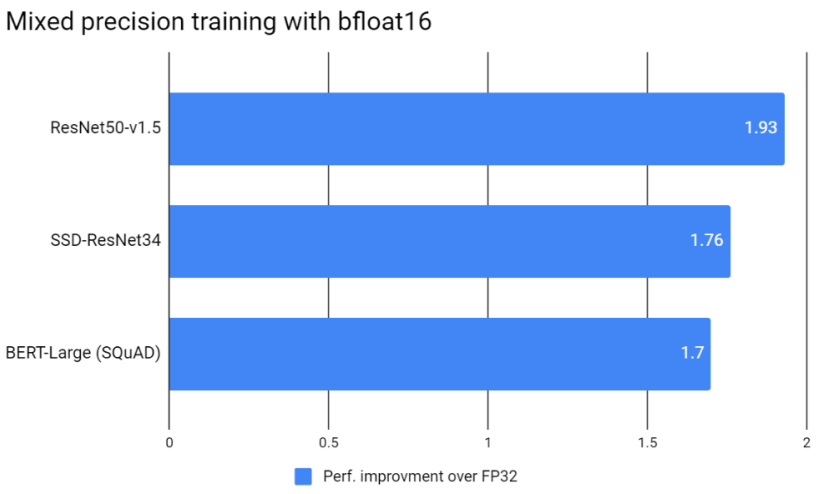
- It can be seen as a short version of FP32, skipping the least significant 16 bits of mantissa
- FP32, and therefore also BF16, offer more than enough range for deep learning training tasks
- FP32 accumulation after the multiply is essential to achieve sufficient numerical behavior on an application level

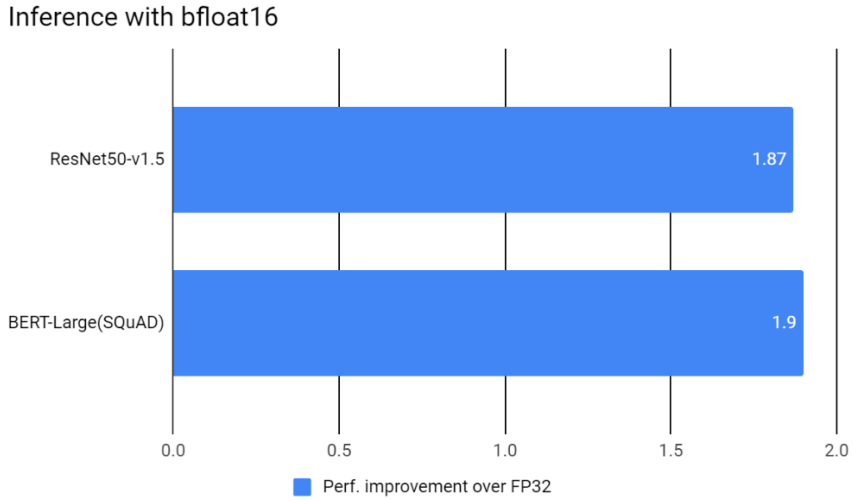


Intel® Deep Learning Boost (Intel® DL Boost) uses bfloat16 format (BF16) on 3rd Gen Intel® Xeon®

- Three new bfloat16 instructions
  - converting to and from bfloat16 data type
    - VCVTNE2PS2BF16
    - VCVTNEPS2BF16
  - Performs a dot product of bfloat16 pairs
    - VDPBF16PS

Performance improvements on 3rd Gen Intel® Xeon® Processors with Tensorflow





Blog :
https://blog.tensorflow.org/2020/06/accelerating-ai-performance-on-3rd-gen-processors-with-tensorflow-bfloat16.html

Performance improvements on 3rd Gen Intel® Xeon® Processors with Pytorch

Training	# Cores per instance	# Instances	BF16 (samples/s)	FP32 (samples/s)	Speedup Ratio
DLRM	28	1	99321	71061	**1.40**
ResNet-50	28	4	399	243	**1.64**
ResNeXt-101 32x4d	28	4	193	120	**1.60**

Inference	# Cores per instance	# Instances	INT8 (samples/s)	FP32 (samples/s)	Speedup Ratio
DLRM	1	28	611082	214559	**2.85**

Blog :

https://www.intel.com/content/www/us/en/artificial-intelligence/posts/intel-facebook-boost-bfloat16.html

Demo : Performance benefit by using Bfloat16 Acceleration on 3rd Gen Intel® Xeon® Processors

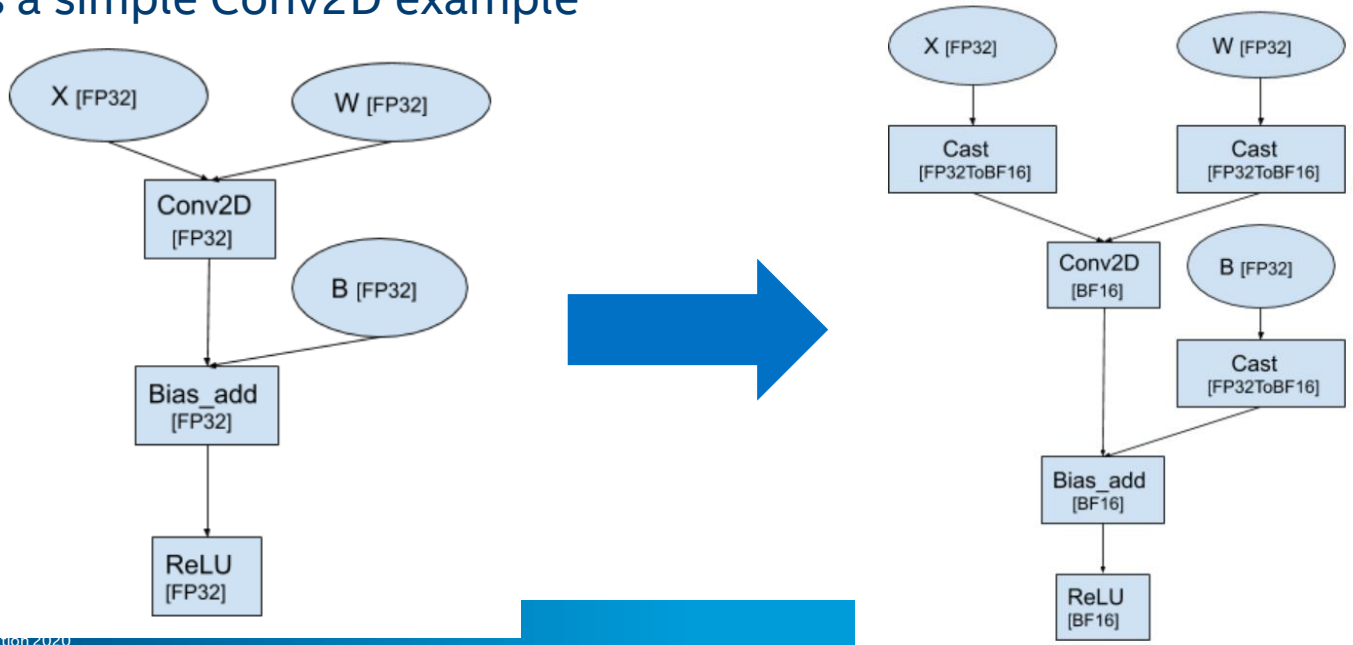
Profile resnet50v1.5 model among different data type

- Training:
  - FP32 vs BF16
- Inference
  - FP32 vs Int8 (quantization required)
  - FP32 vs BF16 (no quantization required)

Model Conversion from FP32 to Bfloat16

AutoMixedPrecisionMkl

- AutoMixedPrecisionMkl is a grappler pass that automatically converts a model written in FP32 data type to operate in BFloat16 data type.
- Here is a simple Conv2D example



Demo : Model conversion from FP32 to Bfloat16

- Convert the graph to BFloat16 on-the-fly as you train the model
- Convert a pre-trained fp32 model to BFloat16

More Details :

https://software.intel.com/content/www/us/en/develop/articles/getting-started-with-automixedprecisionmkl.html

Getting Started with Intel® AI Analytics Toolkit

OVERVIEW

- Visit the AI Kit website for more details and up-to-date product information
- Release Notes

INSTALLATION

- Download the toolkit from Intel, or Anaconda or any of your favorite package managers
- Get started quickly with the AI Kit Docker Container
- Installation Guide
- Utilize the Getting Started Guide

HANDS ON

- Code Samples
- Build, test and remotely run workloads on the Intel® DevCloud for free. No software downloads. No configuration steps. No installations.

LEARNING

- Intel Medium channel for Machine Learning & Analytics Blogs
- Intel AI Blog site
- Webinars and Articles on Tech Decoded

SUPPORT

- Ask questions and share information with others through the Community Forum
- Discuss with experts at AI Frameworks Forum

Q&A

Legal Disclaimer & Optimization Notice <w/o 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 www.intel.com/benchmarks.

INFORMATION IN THIS DOCUMENT IS PROVIDED "AS IS". NO LICENSE, EXPRESS OR IMPLIED, BY ESTOPPEL OR OTHERWISE, TO ANY INTELLECTUAL PROPERTY RIGHTS IS GRANTED BY THIS DOCUMENT. INTEL ASSUMES NO LIABILITY WHATSOEVER AND INTEL DISCLAIMS ANY EXPRESS OR IMPLIED WARRANTY, RELATING TO THIS INFORMATION INCLUDING LIABILITY OR WARRANTIES RELATING TO FITNESS FOR A PARTICULAR PURPOSE, MERCHANTABILITY, OR INFRINGEMENT OF ANY PATENT, COPYRIGHT OR OTHER INTELLECTUAL PROPERTY RIGHT.

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.

Notice revision #20110804

TECH.
DECODED

intel
Software