Solution Brief

Internet of Things

Accelerate TensorFlow on Intel® Architecture with Minimal Code Changes

The OpenVINO™ integration with TensorFlow enables developers to speed up the TensorFlow workflow by adding just two lines of code. They can enhance performance on Intel platforms while using their familiar TensorFlow APIs.

Solution Benefits

- **Fast performance.** The OpenVINO™ integration with TensorFlow enables fast performance for supported models on Intel® architecture.
- **Consistent accuracy.** Accuracy is nearly identical to the original model.
- **Simplicity.** Developers can continue to use the TensorFlow APIs and workflow, with minimal code changes required.

Intel Author:

Kumar Vishwesh,
Product Manager, OpenVINO™ integration with TensorFlow

Executive Overview

The OpenVINO™ toolkit provides a deep learning inference engine that is optimized for Intel processors and accelerators. TensorFlow developers can import their trained models into the toolkit's Model Optimizer and run them through the inference engine to improve performance. However, the process requires code changes and switching out of the TensorFlow environment to optimize the model.

Now, the OpenVINO™ integration with TensorFlow provides in-line performance optimization. Developers can add two lines to their code to import and configure the integration, and it will automatically optimize their models and run them through the OpenVINO™ Inference Engine. The integration supports more than 270 models, but the inference engine falls back to the native TensorFlow engine where operators are not supported. That means anything that works in TensorFlow will work with the OpenVINO™ integration enabled. Speed improvements will depend on the extent to which the model and its operators are supported by the OpenVINO™ integration.

Extreme Vision tested various deep learning models and saw significant geomean acceleration using OpenVINO™ integration with TensorFlow[1]*. The Broad Institute has used the integration to accelerate its genome sequencing workload by 21 percent on an Intel® Core™ i5 processor and by 16.8 percent on an Intel® Xeon® processor[2]*.

Business Challenge: Accelerating TensorFlow Inference

When people think of machine learning performance, they often think of real-time applications, such as safety systems. Saving a fraction of a second could save a life. Speed is also important for more complex inferencing operations, though. A genome sequencing workload, for example, could take days to run. Reducing the time taken could save money in compute costs (especially in the cloud), increase throughput, and ultimately cut waiting times for patients.

The OpenVINO™ toolkit can speed up inferencing performance on Intel® architecture. The toolkit is optimized to use the available processor features, such as Vector Neural Network Instructions (VNNI) in Intel® Xeon® Scalable processors. The toolkit can also be used to enable inferencing on Intel® integrated GPUs, Intel® Movidius™ vision processing units (VPUs), and the Intel® Vision Accelerator Design with 8 Intel® Movidius™ Myriad™ X VPUs.

* See backup for workloads and configurations. Results may vary.

The latter is referred to as VAD-M or HDDL, which is short for High Density Deep Learning. For developers, it's attractive that they don't need to modify their code for the platform it will run on. The same code will run across a range of CPUs and accelerators.

To use the OpenVINO™ toolkit, developers must import their trained model into the OpenVINO™ toolkit Model Optimizer. The Model Optimizer creates an Intermediate Representation (IR), which the OpenVINO™ Inference Engine can process.

Deep learning developers using TensorFlow can import their models into the Model Optimizer and use them in the OpenVINO™ Inference Engine. However, the workflow involves breaking out of the TensorFlow environment, and refactoring the code to use the OpenVINO™ toolkit API.

Solution Value: Inline Optimization for TensorFlow

The OpenVINO™ integration with TensorFlow enables developers to use OpenVINO™ without leaving the TensorFlow environment, and with minimal code changes. Just two lines of code are required to import the OpenVINO™ integration and specify the target processor that will be used (such as CPU, GPU, or Intel Movidius Myriad VPU). Developers can continue to use TensorFlow APIs and do not need to manage the model conversion. It happens automatically in the background.

The OpenVINO™ integration with TensorFlow supports more than 270 TensorFlow models. These include computer vision, natural language processing, and popular TFHub models such as MASK R-CNN, BERT, and GPT2.

Minimal incremental memory and disk footprint are required for the OpenVINO™ integration with TensorFlow. Intel® Core™ and Intel® Xeon® processors are supported, as well as Intel's VPUs and integrated GPUs. Developers can use the OpenVINO™ integration with TensorFlow in a variety of environments, including in the cloud and at the edge, as long as the underlying hardware is based on an Intel® platform.

The solution accelerates performance on TensorFlow while accuracy is nearly identical to the original model:

- Extreme Vision's CV MART is a dedicated cloud for AI. Extreme Vision tested various models for classification, object detection, instance segmentation, and 3D face reconstruction. The company saw significant geomean acceleration using the OpenVINO™ integration with TensorFlow[1]*.
- The Broad Institute worked with Intel to test the performance of the Genome Analysis Toolkit (GATK) using the OpenVINO™ integration with TensorFlow. The results showed a speed-up of 21 percent on an Intel Core i5 processor and 16.8 percent on an Intel Xeon processor[2]*.

The OpenVINO™ integration with TensorFlow is designed for developers who want to use the OpenVINO™ toolkit to enhance inferencing performance with minimal code modifications. For maximum performance, efficiency, tooling customization, and hardware control, we recommend adopting the full Intel® Distribution of OpenVINO™ toolkit, using its native OpenVINO™ APIs and the native OpenVINO™ runtime. See Figure 1 for a breakdown of the differences between the two developer workflows.

* See backup for workloads and configurations. Results may vary.

Criteria	OpenVINO™ toolkit (with Model Optimizer)	OpenVINO™ integration with TensorFlow
Performance	Faster	Slower (model-dependent)
Accuracy	Same	
Ease of use	Needs model conversion	No explicit model conversion required
Model conversion	Offline	Inline
Intel Hardware	All	No support for Field Programmable Gate Arrays (FPGAs)
Compact memory footprint	Yes	Bigger. Needs TensorFlow runtime.
FP16 and INT8 quantization	Yes	Yes
INT4, INT2, Sparsity	Yes	No
OpenVINO™ Deep Learning workbench and other developer tools	Available	Not Available

Figure 1. Comparing the developer workflows using the OpenVINO™ toolkit and the OpenVINO™ integration with TensorFlow.

Solution Architecture: OpenVINO™ integration with TensorFlow

Figure 2 shows the architecture of the OpenVINO™ integration with TensorFlow. The OpenVINO™ integration with TensorFlow includes plug-ins for optimized performance using CPU, GPU, VPU, or HDDL backends. The solution includes the Intel® Math Kernel Library for Deep Neural Networks (Intel® MKL-DNN), which is designed to accelerate deep learning on Intel architecture. It also includes the Compute Library for Deep Neural Networks (clDNN), which helps to accelerate convolutional neural networks (CNNs) on Intel® Processor Graphics.

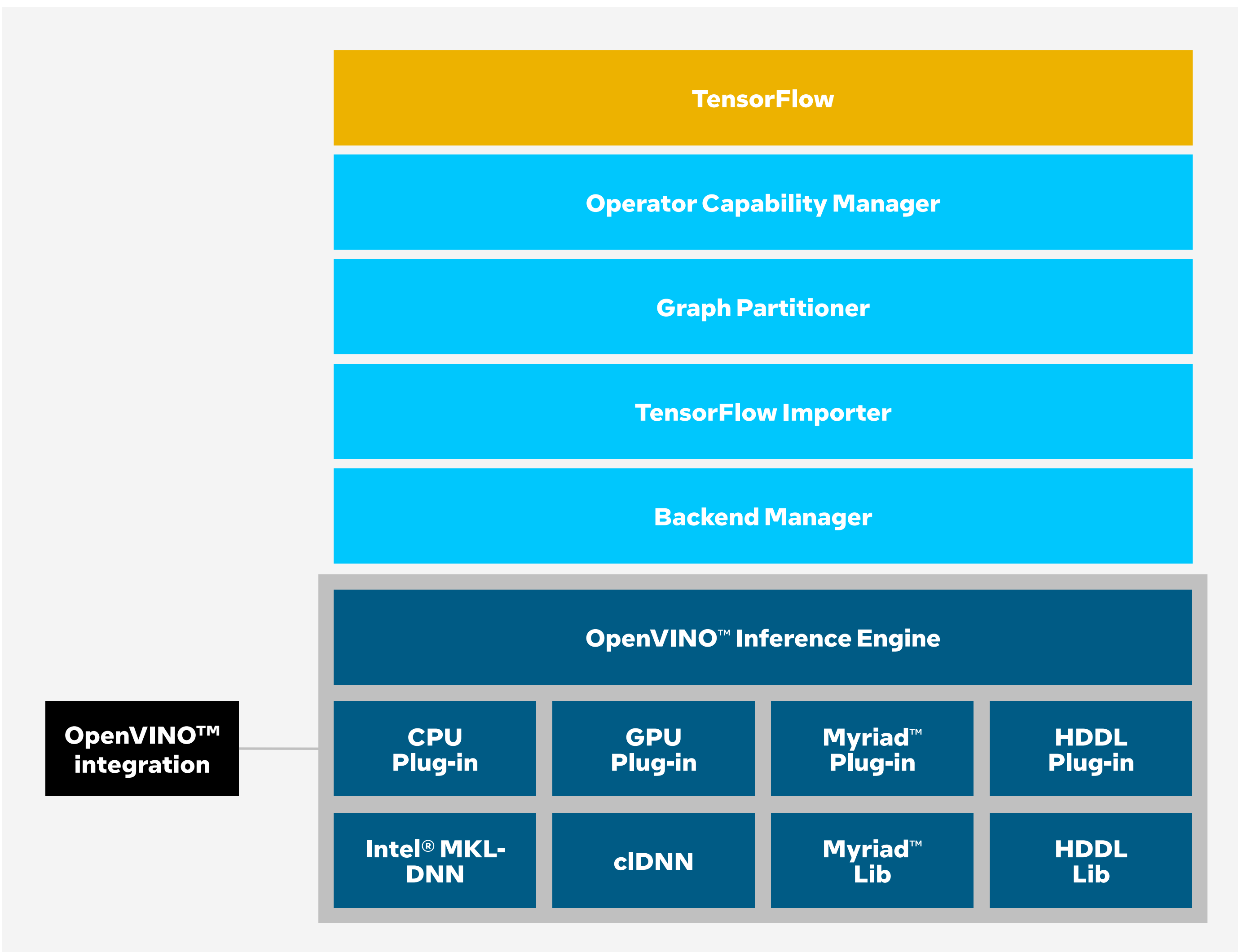

Figure 2. The architecture for the OpenVINO™ integration with TensorFlow.

As shown in Figure 3, the OpenVINO™ integration with TensorFlow falls back to the TensorFlow Inference Engine if the OpenVINO™ Inference Engine is not able to support some or all of the model. For that reason, any model that runs in native TensorFlow can run with the OpenVINO™ integration with TensorFlow enabled. The performance improvement depends on the extent to which the model's operators are supported by the OpenVINO™ integration with TensorFlow.

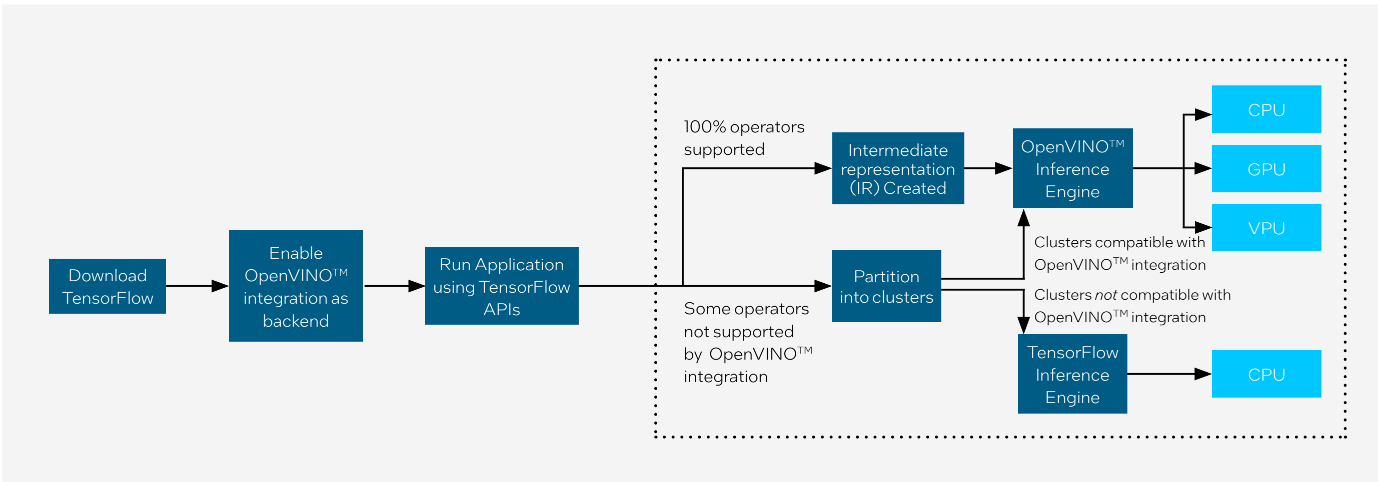

Figure 3. The OpenVINO™ integration with TensorFlow has dynamic fallback for unsupported operators.

Conclusion

Using the OpenVINO™ integration with TensorFlow, developers can easily accelerate supported models on Intel architecture. The solution enables developers to work with their familiar environments and APIs, and use just two lines of code to import and configure the integration.

Developers looking to achieve even higher performance and greater flexibility are recommended to try the Intel Distribution of OpenVINO™ toolkit, using its API and runtime.

Learn More

- **OpenVINO™ integration with TensorFlow at GitHub**
- **OpenVINO™ toolkit documentation**
- **Extreme Vision accelerates TensorFlow workloads using OpenVINO™ integration with TensorFlow**
- **Accelerating Genome Workloads Using the OpenVINO™ Integration with TensorFlow**
- **Using OpenVINO™ to accelerate machine learning tools**
- **Harnessing the power of the Intel® processor's integrated graphics (iGPU)**
- **Validated models supported by OpenVINO™ integration with TensorFlow**

Find the solution that is right for your organization. Contact your Intel representative or register at **Intel IT Center**.

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex.

[1] Configurations: Intel® Core™ i7-9700K, 16GB memory, 300GB SSD, Ubuntu 18.04, TensorFlow version 2.4.1, OpenVINO™ integration with TensorFlow Version 0.5.0, Testing carried out by Extreme Vision. Full details of testing available here: https://bbs.cvmart.net/articles/5412

[2] Configurations: Genome Analysis Toolkit (GATK) running on four-core Intel® Core i5 processor with 8GB RAM. Testing by Intel and Broad Institute. See https://terra.bio/using-openvino-to-accelerate-machine-learning-tools/ and https://www.intel.com/content/www/us/en/artificial-intelligence/posts/accelerating-genome-workloads-ovtf.html

0222/RL/CAT/PDF ♻ Please Recycle 349467-001EN.