Blog (/en-us/blog/) / Data Science (/en-us/blog/topics/datascience/)

Intel and Microsoft bring optimizations to deep learning on Azure

(https://www.facebook.com/share.php?u=https%3A%2F%2Fazure.microsoft.com%2Fblog%2Fintel-and-microsoft-bring-optimizations-to-deep-learning-on-azure%2F) (https://twitter.com/share?url=https%3A%2F%2Fazure.microsoft.com%2Fblog%2Fintel-and-microsoft-bring-optimizations-to-deep-learning-on-azure%2F&text=Intel+and+Microsoft+bring+optimizations+to+deep+learning+on+Azure) (https://www.linkedin.com/shareArticle?mini=true&url=https%3A%2F%2Fazure.microsoft.com%2Fblog%2Fintel-and-microsoft-bring-optimizations-to-deep-learning-on-azure%2F)

Posted on March 7, 2019

Gopi Kumar
Principal Program Manager, Microsoft

This post is co-authored with Ravi Panchumarthy and Mattson Thieme from Intel.

We are happy to announce that Microsoft and Intel are partnering to bring optimized deep learning frameworks to Azure. These optimizations are available in a new offering on the Azure marketplace called the Intel Optimized Data Science VM for Linux (Ubuntu). (https://aka.ms/dsvm/intel)

Over the last few years, deep learning has become the state of the art for several machine learning and cognitive applications. Deep learning is a machine learning technique that leverages neural networks with multiple layers of non-linear transformations, so that the system can learn from data and build accurate models for a wide range of machine learning problems. Computer vision, language understanding, and speech recognition are all examples of deep learning at play today. Innovations in deep neural networks in these domains have enabled these algorithms to reach human level performance in vision, speech recognition (https://blogs.microsoft.com/ai/historic-achievement-microsoft-researchers-reach-human-parity-conversational-speech-recognition/) and machine translation (https://blogs.microsoft.com/ai/machine-translation-news-test-set-human-parity/). Advances in this field continually excite data scientists, organizations and media outlets alike. To many organizations and data scientists, doing deep learning well at scale poses challenges due to technical limitations.

Often, default builds of popular deep learning frameworks like TensorFlow are not fully optimized for training and inference on CPU. In response, Intel has open-sourced framework optimizations for Intel® Xeon processors. Now, through partnering with Microsoft, Intel is helping you accelerate your own deep learning workloads on Microsoft Azure with this new marketplace offering.

> *"Microsoft is always looking at ways in which our customers can get the best performance for a wide range of machine learning scenarios on Azure. We are happy to partner with Intel to combine the toolsets from both the companies and offer them in a convenient pre-integrated package on the Azure marketplace for our users"*
>
> – Venky Veeraraghavan, Partner Group Program manager, ML platform team, Microsoft.

Accelerating Deep Learning Workloads on Azure

Built on the top of the popular Data Science Virtual Machine (DSVM) (https://aka.ms/dsvm), this offer adds on new Python environments that contain Intel's optimized versions of TensorFlow and MXNet. These optimizations leverage the Intel® Advanced Vector Extensions 512 (Intel® AVX-512) (https://www.intel.com/content/www/us/en/architecture-and-technology/avx-512-overview.html) and Intel® Math Kernel Library for Deep Neural Networks (Intel® MKL-DNN) (https://01.org/mkl-dnn) to accelerate training and inference on Intel® Xeon® Processors. When running on an Azure F72s_v2 VM instance, these optimizations yielded an average of 7.7X speedup in training (https://www.intel.ai/intel-optimized-data-science-virtual-machine-azure) throughput across all standard CNN topologies. You can find more details on the optimization practice here (https://www.intel.ai/intel-optimized-data-science-virtual-machine-azure).

As a data scientist or AI developer, this change is quite transparent. You still code with the standard TensorFlow or MXNet frameworks. You can also use the new set of Python (conda) environments (intel_tensorflow_p36, intel_mxnet_p36) on the DSVM to run your code to take full advantage of all the optimizations on an Intel® Xeon Processor based F-Series or H-Series VM instance on Azure. Since this product is built using the DSVM as the base image, all the rich tools for data science and machine learning (https://docs.microsoft.com/azure/machine-learning/data-science-virtual-machine/overview#whats-included-in-the-data-science-vm) are still available to you. Once you develop your code and train your models, you can deploy them for inferencing on either the cloud or edge.

> *"Intel and Microsoft are committed to democratizing artificial intelligence by making it easy for developers and data scientists to take advantage of Intel hardware and software optimizations on Azure for machine learning applications. The Intel Optimized Data Science Virtual Machine (DSVM) provides up to a 7.7X speedup on existing frameworks without code modifications, benefiting Microsoft and Intel customers worldwide"*
>
> – Binay Ackalloor, Director Business Development, AI Products Group, Intel.

Performance

In Intel's benchmark tests run on Azure F72s_v2 instance, here are the results comparing the optimized version of TensorFlow with the standard TensorFlow builds.

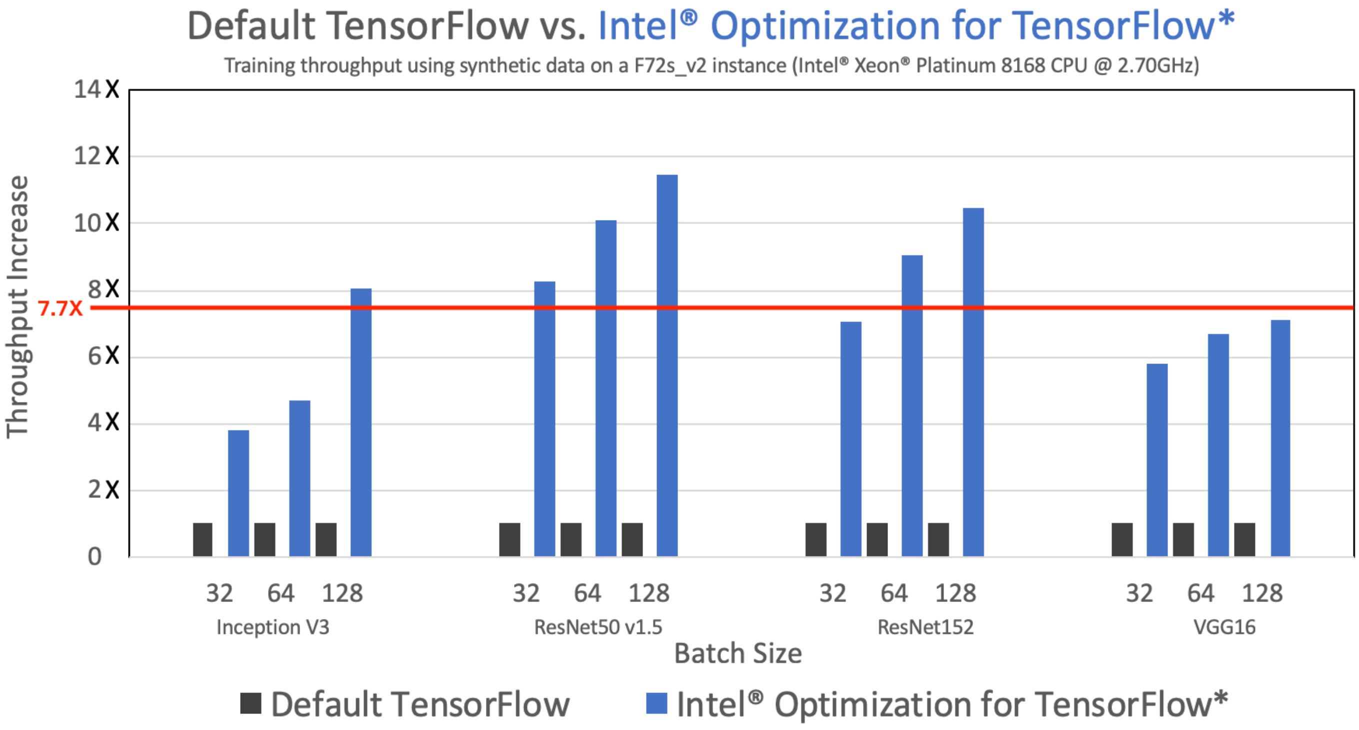
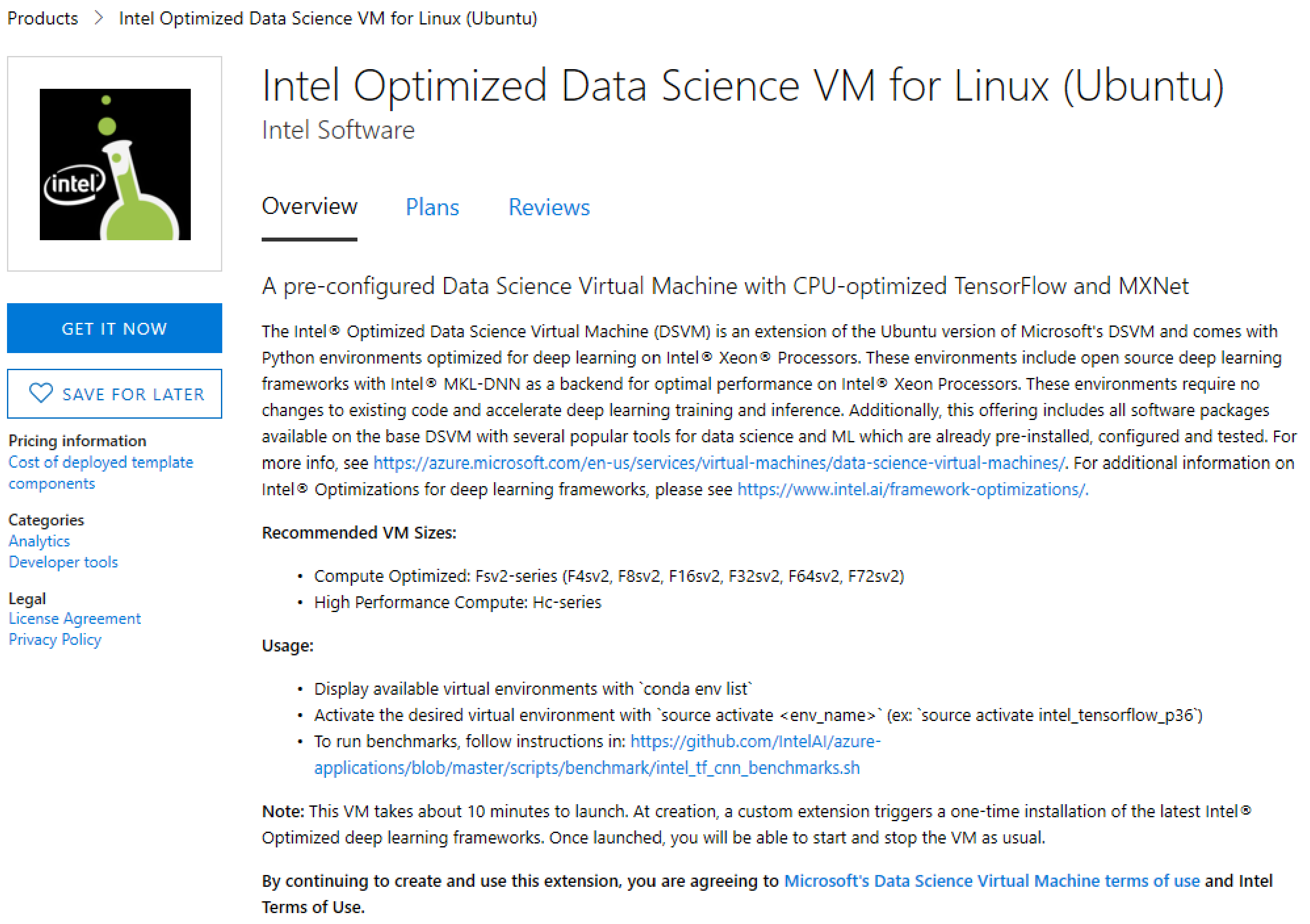


Figure 1: Intel® Optimization for TensorFlow provides an **average of 7.7X increase** (average indicated by the red line) in training throughput on major CNN topologies. Run your own benchmarks using tf_cnn_benchmarks (https://github.com/IntelAI/azure-applications/tree/master/scripts/benchmark). Performance results are based on Intel testing as of 01/15/2019. Find the complete testing configuration here (https://www.intel.ai/intel-optimized-data-science-virtual-machine-azure).

Getting Started

To get started with the Intel Optimized DSVM, click on the offer in the Azure Marketplace (https://aka.ms/dsvm/intel), then click "GET IT NOW". Once you answer a few simple questions on the Azure Portal, your VM is created with all the DSVM tool sets and the Intel optimized deep learning frameworks pre-configured and ready to use.


Products > Intel Optimized Data Science VM for Linux (Ubuntu)
Overview
Plans
Reviews
GET IT NOW
SAVE FOR LATER

Intel Optimized Data Science VM for Linux (Ubuntu)

Intel Software

Pricing information
Cost of deployed template components

Categories
Analytics
Developer tools

Legal
License Agreement
Privacy Policy

A pre-configured Data Science Virtual Machine with CPU-optimized TensorFlow and MXNet

The Intel® Optimized Data Science Virtual Machine (DSVM) is an extension of the Ubuntu version of Microsoft's DSVM and comes with Python environments optimized for deep learning on Intel® Xeon® Processors. These environments include open source deep learning frameworks with Intel® MKL-DNN as a backend for optimal performance on Intel® Xeon Processors. These environments require no changes to existing code and accelerate deep learning training and inference. Additionally, this offering includes all software packages available on the base DSVM with several popular tools for data science and ML which are already pre-installed, configured and tested. For more info, see https://azure.microsoft.com/en-us/services/virtual-machines/data-science-virtual-machines/. For additional information on Intel® Optimizations for deep learning frameworks, please see https://www.intel.ai/framework-optimizations/.

Recommended VM Sizes:

- Compute Optimized: Fsv2-series (F4sv2, F8sv2, F16sv2, F32sv2, F64sv2, F72sv2)
- High Performance Compute: Hc-series

Usage:

- Display available virtual environments with `conda env list`
- Activate the desired virtual environment with `source activate <env_name>` (ex: `source activate intel_tensorflow_p36`)
- To run benchmarks, follow instructions in: https://github.com/IntelAI/azure-applications/blob/master/scripts/benchmark/intel_tf_cnn_benchmarks.sh

Note: This VM takes about 10 minutes to launch. At creation, a custom extension triggers a one-time installation of the latest Intel® Optimized deep learning frameworks. Once launched, you will be able to start and stop the VM as usual.

By continuing to create and use this extension, you are agreeing to Microsoft's Data Science Virtual Machine terms of use and Intel Terms of Use.

The Intel Optimized Data Science VM is an ideal environment to develop and train deep learning models on Intel Xeon processor-based VM instances on Azure. Microsoft and Intel will continue their long partnership to explore additional AI solutions and framework optimizations to other services on Azure like the Azure Machine Learning service (https://aka.ms/azureml) and Azure IoT Edge (https://azure.microsoft.com/services/iot-edge/).

Next steps

- Create your Intel Optimized Data Science VM (https://aka.ms/dsvm/intel) instance from the Azure Marketplace.
- Learn more about the Intel Optimized Data Science VM (https://www.intel.ai/intel-optimized-data-science-virtual-machine-azure).
- Build AI solutions and deploy machine learning models in production at scale using Azure Machine Learning service (https://azure.microsoft.com/services/machine-learning-service/).

- New to Azure? Get your free trial (https://azure.com/free).

Data Science (/en-us/blog/topics/datascience/) Artificial Intelligence (/en-us/blog/topics/artificial-intelligence/) Machine Learning (/en-us/blog/topics/machine-learning-2/) DSVM (/en-us/blog/tag/dsvm/) Deep Learning (/en-us/blog/tag/deep-learning/) Tensorflow (/en-us/blog/tag/tensorflow/) MXNet (/en-us/blog/tag/mxnet/) Intel (/en-us/blog/tag/intel/)

Subscribe (//azurecomcdn.azureedge.net/en-us/blog/feed/)

Explore

See where we're heading. Check out upcoming changes to Azure products

Azure updates (/en-us/updates/)

Let us know what you think of Azure and what you would like to see in the future

Provide feedback (https://feedback.azure.com)

Topics

Announcements (/en-us/blog/topics/announcements/) (2088)

API Management (/en-us/blog/topics/api-management-2/) (31)

Artificial Intelligence (/en-us/blog/topics/artificial-intelligence/) (195)

Azure Maps (/en-us/blog/topics/azure-maps/) (16)

Azure Marketplace (/en-us/blog/topics/azure-marketplace/) (135)

Azure Stream Analytics (/en-us/blog/topics/azure-stream-analytics/) (29)

Big Data (/en-us/blog/topics/big-data/) (623)

Blockchain (/en-us/blog/topics/blockchain/) (84)

Business Intelligence (/en-us/blog/topics/business-intelligence/) (113)

Cloud Strategy (/en-us/blog/topics/cloud-strategy/) (599)

Cognitive Services (/en-us/blog/topics/cognitive-services/) (114)

Data Science (/en-us/blog/topics/datascience/) (102)

Data Warehouse (/en-us/blog/topics/data-warehouse/) (204)

Database (/en-us/blog/topics/database/) (576)

Developer (/en-us/blog/topics/developer/) (1145)

DevOps (/en-us/blog/topics/devops/) (65)

Events (/en-us/blog/topics/events/) (214)

Government (/en-us/blog/topics/government/) (59)

Hybrid (/en-us/blog/topics/hybrid/) (49)

Identity & Access Management (/en-us/blog/topics/identity-access-management/) (85)

Internet of Things (/en-us/blog/topics/internet-of-things/) (321)

IT Pro (/en-us/blog/topics/it-pro/) (580)

Last week in Azure (/en-us/blog/topics/last-week-in-azure/) (92)

Machine Learning (/en-us/blog/topics/machine-learning-2/) (18)

Management (/en-us/blog/topics/management/) (309)

Media Services & CDN (/en-us/blog/topics/media-services/) (202)

Mobile (/en-us/blog/topics/mobile/) (154)

Monitoring (/en-us/blog/topics/monitor/) (118)

Networking (/en-us/blog/topics/networking/) (201)

Partner (/en-us/blog/topics/partner/) (89)

Security (/en-us/blog/topics/security/) (353)

Serverless (/en-us/blog/topics/serverless/) (61)

Storage, Backup & Recovery (/en-us/blog/topics/storage-backup-and-recovery/) (641)

Supportability (/en-us/blog/topics/supportability/) (41)

Updates (/en-us/blog/topics/updates/) (534)

Virtual Machines (/en-us/blog/topics/virtual-machines/) (646)

Web (/en-us/blog/topics/web/) (359)

Articles by date

October 2019 (/en-us/blog/2019/10/)

September 2019 (/en-us/blog/2019/09/)

August 2019 (/en-us/blog/2019/08/)

July 2019 (/en-us/blog/2019/07/)

June 2019 (/en-us/blog/2019/06/)

May 2019 (/en-us/blog/2019/05/)

Full archive (/en-us/blog/archives/)