Solution Brief

Software Development
AI/ML Development/Deployment

Improve MLOps and Accelerate Model Deployment with Paperspace and Intel

Paperspace's Gradient platform optimizes machine learning pipelines and delivers faster inferencing and lower query latency on 2nd Gen Intel® Xeon® processors using Intel Distribution of OpenVINO™ toolkit and Intel Distribution of OpenVINO Model Server

Intel© AI Builders Member

gradient by Paperspace

"We have worked closely with Intel AI Builders to help optimize Gradient for Xeon processors and to create containers that natively support instruction sets on the latest Intel Xeon Scalable processors. The Paperspace team has greatly benefited by integrating tools like OpenVINO, and making it easier for people to deploy to Xeon processors in the data center."

Dillon Erb, Founder & CEO, Paperspace

As DevOps did a decade ago with software teams, Machine Learning Ops (MLOps) helps improve efficiency in Artificial Intelligence (AI) model development, training, and deployment. Training, maintaining, and deploying an AI model can be a very long and complex process with a lot of moving parts. Building and maintaining an ML pipeline to support MLOps is time consuming; teams often spend more time on tooling and infrastructure than on training models. Further, without orchestration software, an entire ML pipeline is extremely difficult to manage. All of this becomes more challenging and time consuming as the amount of data and algorithms scale.

MLOps infrastructure solutions are available as both cloud services and on-premise software to help development teams adopt an ML pipeline quickly. Alternatively, IT departments can build their own.

Enterprises integrating AI into their product lines will need to make the MLOps solution subscribe-versus-buy-versus-build decision. In that process, it is important to ensure that the choice of pipeline supports all of an MLOps team's languages, approaches, frameworks, and technologies employed. This is especially true as new technologies emerge and evolve. Technologies, such as the OpenVINO™ toolkit and Intel® AI software stack to support Intel XPUs can deliver improved performance to company solutions. When establishing an MLOps pipeline, a team should be able to accommodate innovations such as these as they emerge.

In the end, a seamless and simple way of creating an ML pipeline that supports all the needs of MLOps is critical for a successful AI implementation.

Paperspace offers an MLOps pipeline in the cloud. Their Gradient platform exposes an API which allows for quick training and deployment of AI models on various cloud instance types. Their tools and offerings support the latest technologies and software from Intel to help developers deliver powerful AI/ML solutions.

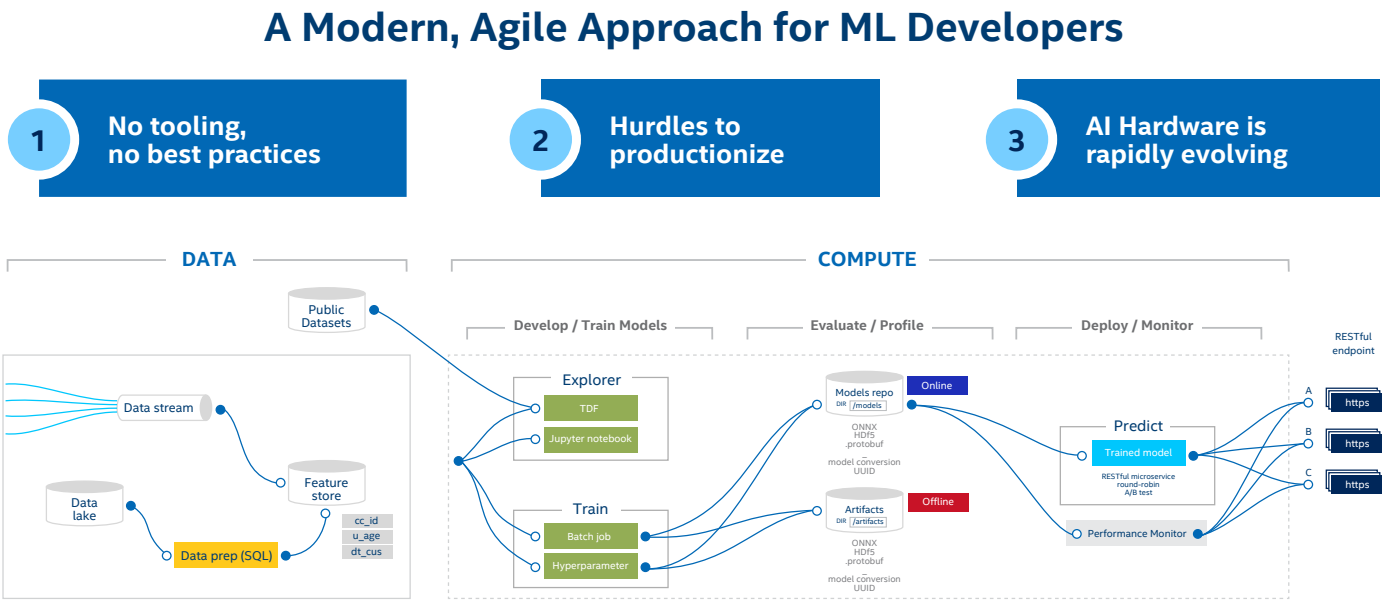

Figure 1. The Gradient software stack enables fast ML development across many frameworks and technologies.

MLOps With Paperspace Gradient

Paperspace Gradient is an MLOps platform for software development teams that produce machine learning models. The platform provides tooling, infrastructure, and lifecycle management to help teams build and deploy models with continuous integration/continuous deployment (CI/CD) best practices (Figure 1).

Gradient provides abstraction layers across the entire pipeline—including notebooks, datasets, experiments, models, and clusters. It is available as a web application, command line interface (CLI), and for enterprise clients to self-host as a Kubernetes application. Figure 2 shows an example of the interface.

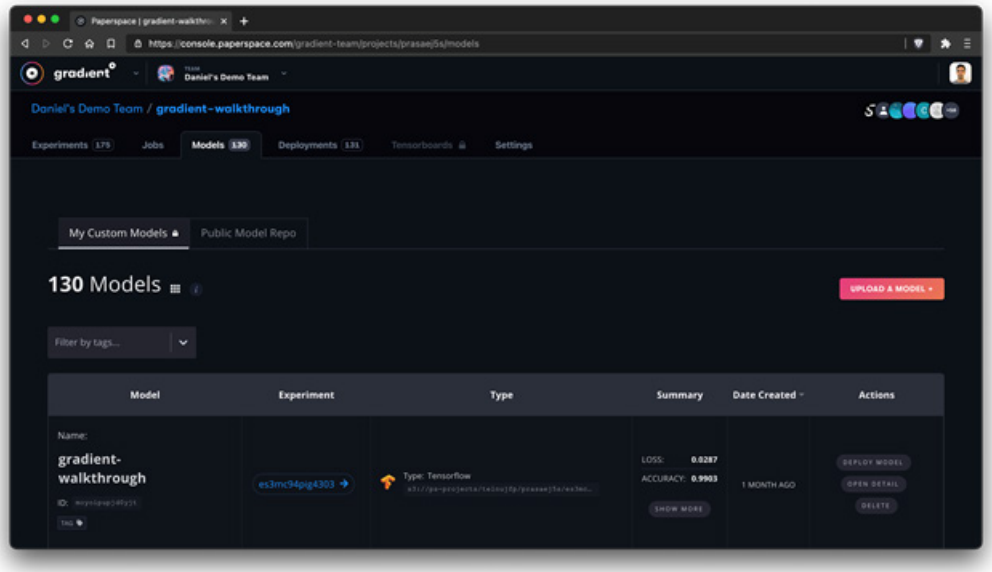

Figure 2. Gradient's cloud-based solution simplifies access to a powerful infrastructure for MLOps.

Accelerating Inferencing on Gradient with Intel Optimizations

Gradient supports MLOps teams that use different workflows, frameworks, and hardware to achieve their development and deployment goals. The Intel OpenVINO toolkit allows computer vision model inferencing across a wide range of hardware rather than requiring companies to deploy on platforms around which the model was originally developed. Those platforms might be limited in availability and expensive to use for inferencing. OpenVINO enables great flexibility in deployment across different hardware options that meet the needs of the application at hand. Thus, Paperspace supports the OpenVINO toolkit and the Intel Distribution of OpenVINO toolkit for inferencing on Intel XPUs.

Gradient Integrates the OpenVINO Model Server for Inferencing

As part of the OpenVINO project, the OpenVINO Model Server (OVMS) is a scalable, high-performance solution for serving machine learning models optimized for Intel architectures. The server provides an inference service via gRPC endpoint or REST API. OpenVINO Model Server makes it easy to deploy new algorithms and AI experiments using the same architecture as TensorFlow Serving for any models trained in a framework that is supported by OpenVINO toolkit.[1]

By leveraging OpenVINO model serving in their Gradient pipeline and optimizing for Intel architecture, Paperspace showed that customers are able to improve inference performance on 2nd Gen Intel Xeon® Scalable processors by up to 1.6X.[2]

Accelerating Inferencing Up to 1.6X[2] with Intel Optimizations

Inferencing speedup was achieved by first leveraging the OpenVINO model optimizer tool to convert existing TensorFlow/ONNX models to run more efficiently on 2nd Gen Intel Xeon Scalable processors. The Model Optimizer is a cross-platform command-line tool that facilitates the transition between the training and deployment environment, performs static model analysis, and adjusts deep learning models for optimal execution on end-point target devices.[3]

Training of models using the COCO dataset was accomplished using a Docker container optimized for Intel architecture. The container incorporated the Intel Math Kernel Library for Deep Neural Networks (Intel MKL-DNN) into the standard Horovod-CPU Docker file definition available in Gradient. The Intel MKL-DNN uses Intel Advanced Vector Extensions 512 (Intel AVX512) for accelerated floating-point processing.

Inference tests were run on Intel Xeon Gold 6230R processor-based servers available in the Gradient platform. All configurations were identical except for the difference in the models.

- Standard TensorFlow model on the COCO dataset
- OpenVINO model with Intel Distribution of OpenVINO toolkit and Intel Distribution of OpenVINO Model Server with the COCO dataset

The OpenVINO Model Server was chosen to increase throughput and reduce the prediction latency of requests. The results are shown in Figure 3.

With these optimizations, results return quicker.

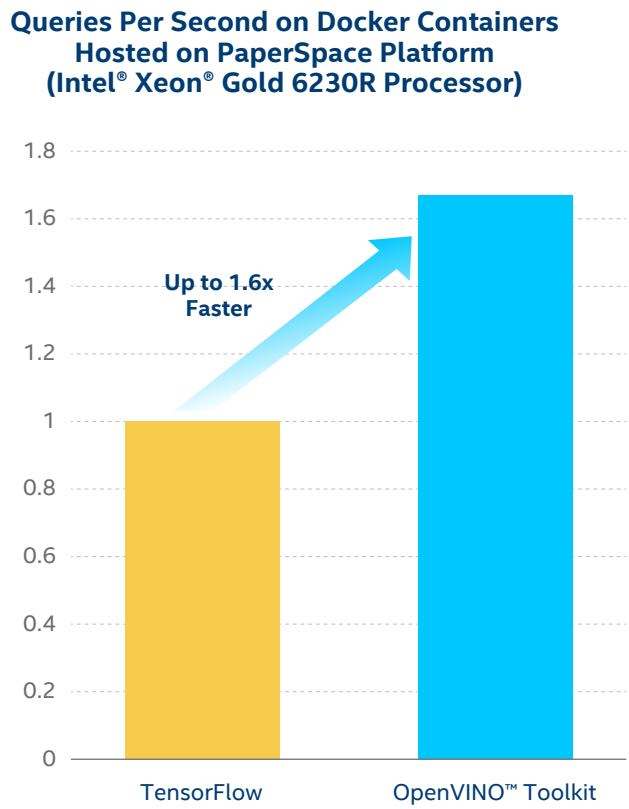

Figure 3. Inferencing speedup with Intel® Distribution of OpenVINO™ toolkit.[2]

Faster inferencing on 2nd Gen Intel Xeon processors means less time on the server, which can reduce costs to end users on a per-request basis.

Conclusion

Developing and managing models more efficiently and running models faster delivers results and insights quicker to users and applications, which can reduce costs. The Paperspace Gradient MLOps pipeline solution allows model-building teams to work more efficiently with proven tools that will support new technologies as they emerge. A recent innovation, OpenVINO allows running inferencing on a wide range of Intel architectures. The Gradient MLOps platform supports Intel AI software and technologies. With TensorFlow models using OpenVINO and optimized for both the OpenVINO toolkit and OpenVINO Model Server, inferencing runs up to 1.6X faster with lower latency of requests, as illustrated by the above testing. Optimizations and tools that enable both a more efficient development pipeline and faster inferencing on Intel architecture offer value to both the teams and the inference application.

For more information about Paperspace, visit **paperspace.com**

Find out how the Gradient platform can help your MLOps teams at **gradient.paperspace.com**

To learn more about running Gradient on Intel Xeon infrastructure, visit **gradient.paperspace.com/intel-ai**

Learn more about the Intel AI Builders program at **builders.intel.com/ai**

Intel® AI Builders Member

Paperspace is an Infrastructure as a Service (IaaS) company and provider of AI tools for developers. They facilitate the use of CPUs for AI projects with a software layer that automates provisioning and managing of that infrastructure.

[1] https://software.intel.com/content/www/us/en/develop/articles/containers/openvino-model-server.html.

[2] Testing of model performance optimized with Intel Distribution of OpenVINO toolkit and Intel® Distribution of OpenVINO Model Server on 2nd Gen Intel® Xeon® Gold 6230R processors:

NEW: Tested by Intel as of September 16, 2020 on Paperspace platform. 2-socket 2nd Gen Intel® Xeon® Gold 6230R CPU, 26 cores HT On/52 threads, Turbo ON, Total Memory 192 GB, BIOS (including microcode version: cat /proc/cpuinfo | grep microcode –m1): SE5C620.86B.02.01.0011.032620200659, Kernel: 5.3.0-61-generic, Mitigation variants (1,2,3,3a,4, L1TF) https://github.com/speed47/spectre-meltdown-checker: Full Mitigation.

OS: Ubuntu 18.04.4 LTS . Deep Learning Framework: TensorFlow, Intel® Optimizations for TensorFlow (pip install), Intel® Distribution of OpenVINO toolkit and Intel® Distribution of OpenVINO Model Server

Baseline: Same hardware/software/frameworks as above. Models run without OpenVINO implementation.

[3] https://docs.openvinotoolkit.org/2019_R3/_docs_MO_DG_Deep_Learning_Model_Optimizer_DevGuide.html.

Performance results are based on testing as of dates shown in configuration and may not reflect all publicly available security updates. No product can be absolutely secure. See configuration disclosure for details.

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.

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

Printed in USA 0121/AU/HBD/PDF ♻ Please Recycle 345505-001US

intel®