Solution Brief

Education/Career Development
Computer Vision Inferencing

NimbleBox's Intel Optimizations Speed Up In-Cloud Inferencing Performance

Intel® Distribution of OpenVINO™ toolkit and Intel optimizations for machine learning frameworks and languages boost inferencing up to 20x on popular AI models running on Intel CPUs

Intel© AI Builders Member

NimbleBox.ai

Artificial Intelligence (AI) and machine learning (ML) are transforming almost every major industry. Their growing adoption requires giving more developers and data scientists access to hardware, software, frameworks, and a development environment—either in-house or in-cloud. Setting up such an environment on-premises is a complex and expensive process. It can require specialized hardware (such as costly GPUs, a range of processors, and FPGAs) depending on the use case and intended inferencing endpoint device. Increasingly, organizations want to do inferencing on-premises to optimize performance and ease deployment. But costs of GPUs make it infeasible, especially for companies with limited funding. Complicating this situation, with AI/ML in continuous innovation, developers and data scientists need easy access to emerging tools, languages, and frameworks that accelerate AI. These tools require acquisition, installation, and management.

ML and inferencing resources can be set up using cloud services. But, with the many choices of cloud configurations and specific development needs, it is not as simple or speedy a process as a single click to create the right environment. For developers and data scientists to configure their own cloud takes time away from the focus of their expertise. For IT to create specific cloud configurations for different teams or a general solution expands their already strained roles and human resources.

NimbleBox is an AI/ML/inferencing Platform as a Service (PaaS) that allows data science teams, developers, and students to work and collaborate on their deep learning projects with ease. NimbleBox, with its one-click process (Figure 1), lets customers quickly build their own machine learning and inferencing environments. Customers can select resources based on processors only or with GPU accelerators and get to work quickly. The platform offers a wide range of current AI/ML frameworks, languages, and libraries widely used for development and inferencing. With a built-in developer-centric AI workbench, users can quickly create and train models and inference them, whether they work for enterprise, are students beginning their AI journey, or self-learners.

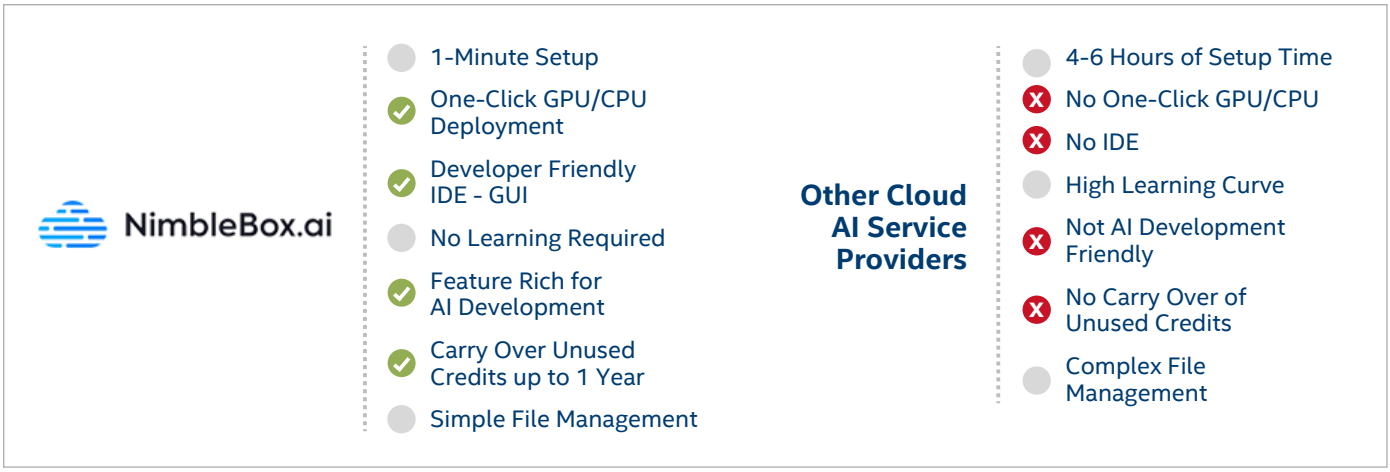

Figure 1. Differences between NimbleBox and generic cloud AI offerings.

NimbleBox Enables Inferencing Performance at Low Cost on Intel xPUs

While industries have seen continued improvements in AI/ML hardware and algorithms, the major bottlenecks remain to be performance and cost. Model training is often demanding on computing resources, but models perform well during inferencing. And, while expensive GPUs are often promoted as high-performance inferencing devices, optimizing algorithms for CPUs can lead to significant performance improvements at low cost. NimbleBox offers solutions that can accelerate inferencing on CPUs using Intel® AI Technologies.

Intel AI technologies are part of a foundational stack of hardware and software that enables and accelerates training and inferencing. Intel processors are ubiquitous across cloud and enterprise data centers and edge computing. Additionally, the Intel Distribution of the OpenVINO™ toolkit provides an approach that allows developers to use a single code base to deploy algorithms for inferencing across a wide range of Intel xPUs. These devices include Intel processors, Intel GPUs, Intel FPGAs, and Intel Movidius™ Neural Compute stick. Such options create deployment flexibility and cost-saving choices without sacrificing performance. They allow companies to optimize cost and performance for their particular inferencing applications.

With Intel Distribution of OpenVINO toolkit and other tools along with Intel AI Technology choices in their platform, NimbleBox offers high-performance, low-cost inferencing capabilities. Their solution is applicable to any type of user or scale of operation, from a budget-conscious student or independent developer to large enterprise development teams. To evaluate performance benefits of their offerings, NimbleBox engineers benchmarked several popular models using Intel AI Technologies and the Intel Distribution of OpenVINO toolkit.

Benchmarking Popular Computer Vision Models

For the benchmarking, NimbleBox used commercial workloads, AI frameworks, and programming environments running on 2nd Generation Intel Xeon® Scalable Processors. They achieved impressive results that enable high-performance inferencing on low-cost Intel architecture.

Engineers inferenced four popular deep learning computer vision models for comparative analysis between those models optimized with Intel Distribution of OpenVINO toolkit and non-optimized versions. The four models included YOLOv3, Mask R-CNN, FaceNET, and OpenPOSE. The benchmarks measured inference time between the non-optimized models that used TensorFlow, PyTorch, and Caffe frameworks alone, and those models that used OpenVINO and optimized with the Intel Distribution of OpenVINO toolkit. They were run on CPU-only (no GPU accelerators) instances with Intel Xeon 8275CL processors.

NimbleBox developers analyzed model latency, throughput, and cost-effectiveness, considering the costs of various compute engines and discrete GPUs. Intel processor performance, cost of hardware and installation, time trade-off, and other parameters were also kept in consideration for deployment on the NimbleBox cloud service and on-premises.

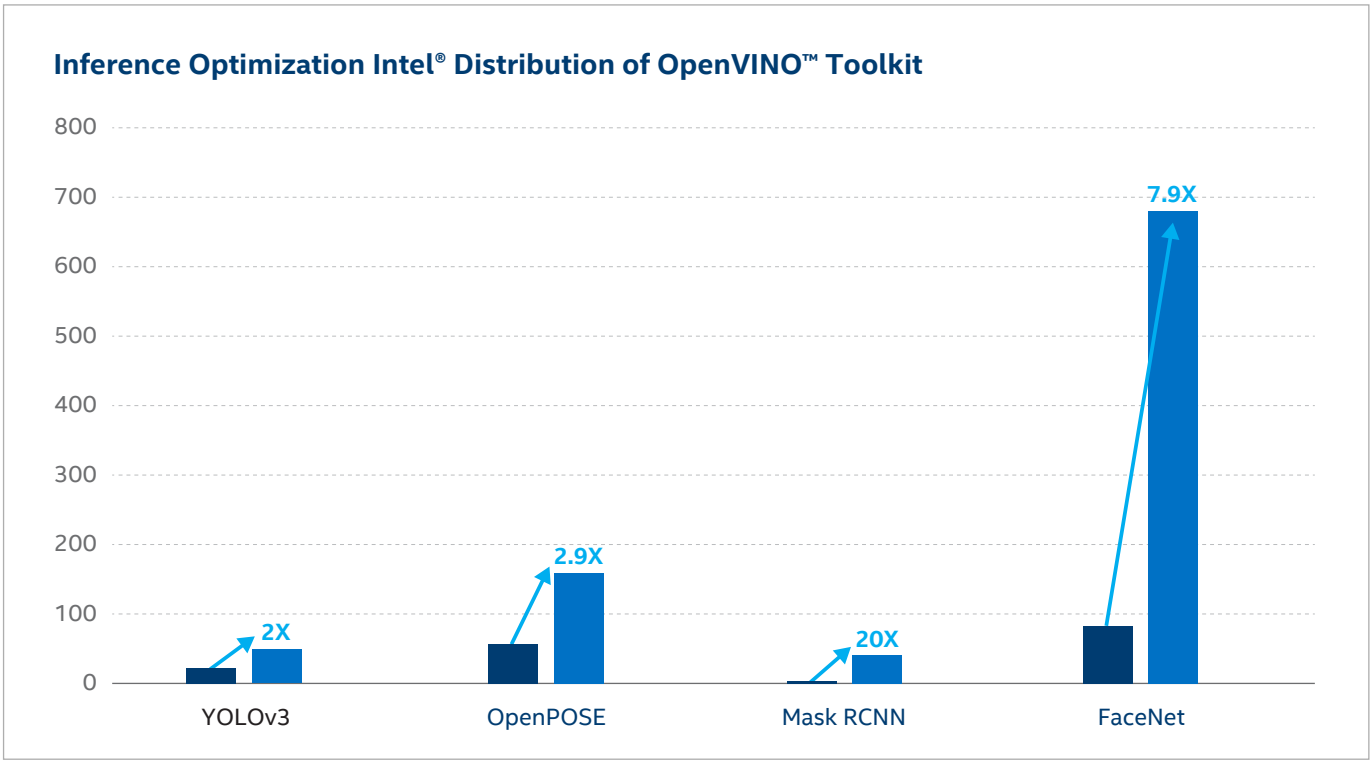

Figure 2. Benchmark results of four computer vision models optimized with Intel Distribution of OpenVINO toolkit.[1]

Up to 20X Performance Boost with Intel Optimizations[1]

NimbleBox developers worked with Intel engineers to gain in-depth knowledge about Intel AI hardware and software offerings and how best to maximize performance using them. The benefits of the benchmarking and understandings from Intel will help NimbleBox customers make informed choices in their AI journey.

Between the non-optimized and optimized models, NimbleBox achieved up to a 20x increase in inference performance over non-optimized versions on IA. The results of the benchmarks across the four models are shown in Figure 2.

This inference workload used a COCO dataset with approximately 100 percent vectorization and achieved about 90 percent CPU utilization. High vectorization means that the code efficiently leveraged the Intel Advanced Vector Extensions 512 and Intel Deep Learning Boost with Vector Neural Network Instructions (VNNI) to boost performance. At high CPU utilization, no compute cycles were wasted. The workload took advantage of nearly all of the CPU resources made available over the lifetime of the job. Since users pay for the instance time, high utilization means high-cost efficiency (more compute/dollar) for the project.

Faster inferencing can deliver insights quicker. With options to run inferencing jobs across different types and cost-levels of IA, users can balance the need for performance and operational budgets of their projects more effectively. They have more choices with which to optimize value at the best attainable performance.

Conclusion

Integrating optimized computer vision models and the Intel AI offerings, NimbleBox provides developers a flexible cloud-based ML development and deployment platform with access to emerging technologies. By including these optimization options, developers can achieve performance gains with powerful, cost-effective IA compute resources. NimbleBox on Intel architecture delivers both performance and value for ML inferencing applications that are impacting nearly every industry.

For more information about the NimbleBox, visit **nimblebox.ai**
Learn more about the Intel AI Builders program at **builders.intel.com/ai**

Intel© AI Builders Member

NimbleBox.ai is a Platform as a Service provider for machine learning, big data and deep learning projects. Their one-click process lets customers quickly build their own machine learning and inferencing environment for developing algorithms and running their workloads.

[1] Model benchmarking of model performance optimized with Intel Distribution of OpenVINO™ toolkit on 2nd Gen Intel® Xeon® Platinum 8275CL processors:

NEW: Tested by NimbleBox as of September 16, 2020 on AWS. 2-socket 2nd Gen Intel® Xeon® Platinum 8275CL CPU @ 3.00 GHz, 24 cores HT On, Turbo ON, Total Memory 92 GB, BIOS: 1.0 (ucode:0x5002f01), Ubuntu 18.04.5 LTS, Linux ip-172-31-5-45 5.3.0-1032-aws #34~18.04.2-Ubuntu SMP Fri Jul 24 10:06:28 UTC 2020 x86_64 GNU/Linux, Deep Learning Framework: TensorFlow 1.15 and 2.0.0, Keras 2.3.0, Intel® Distribution of OpenVINO™ toolkit, Intel® Optimizations for TensorFlow (pip install), Python 3.6.9.

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

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 0221/AU/HBD/PDF ♻ Please Recycle 345406-001US

intel®