SOLUTION BRIEF

AI Model Development
Neural Networks

DarwinAI Generative Synthesis* Platform and Intel® Optimizations for TensorFlow* Accelerate Neural Networks

INTEL® AI BUILDERS MEMBER

DARWIN AI

Up to 16.3X[1] speedup on standard image classification CNNs

By definition, deep neural networks (DNNs) are complex systems, making them difficult to build, run, and describe. This complexity can result in inference times that do not meet the needs of field deployments. To aid data scientists and developers, algorithms can be used to build and evaluate DNNs. DNNs themselves are ideal targets on which to apply artificial intelligence (AI) solutions.

Create Small DNNs with Low Inference Times and Explainability

To simplify complex DNNs and reduce inference times, DarwinAI created its Generative Synthesis platform, which uses AI itself to examine a neural network. The platform builds an understanding of the network and then generates several highly compact, faster versions, which maintain functional fidelity. The system even explains network predictions. This patented "AI building AI" technology:

- Dramatically reduces the size, complexity, and guesswork in designing efficient, high-performance deep learning solutions.
- Maintains functional accuracy and reduces inference times for real world applications of deep neural networks.
- Facilitates "explainable" deep learning—the ability to understand why a network makes the decisions it does—which is particularly important in regulated industries.
- Reduces computational requirements by generating highly optimized models (particularly useful in deploying deep learning at the edge).

Dramatically Enhances Productivity

Today, development of deep learning models can require some guesswork by the data scientist. For example, data scientists must consider what reference model to begin with, how that model performs with certain training sets, and what holes and imbalances may exist in the data set.

The Generative Synthesis platform allows developers to collaborate with state-of-the art AI tools to reduce this guesswork. Instead of tuning a model for weeks or months, the platform generates optimized models in days, based on user-defined requirements. A developer then collaborates with the platform to choose the right network for the task at hand. The platform can generate additional networks based on more detailed requirements.

The result is a dramatic acceleration of deep learning development.

Explainable AI for Deeper Insights

In regulated industries, approvals can require describing—sometimes in detail—how a system works. For products leveraging the power of AI, the manufacturer/developer must be able to describe how the product "thinks."

The DarwinAI Generative Synthesis platform provides granular insights into neural network performance with details about the network's decision-making process for specific tasks at the layer or neuron level. A deeper understanding of the network's processes gives developers greater insight into its internal workings for fine tuning efficiency and accuracy, and it can help meet the requirements of some approval processes.

Generative Synthesis Platform

The Generative Synthesis platform uses TensorFlow* and provides a web-based interface for orchestration. The platform is typically made available by way of Docker* images and is hosted in the client's environment, which is important for enterprise customers that do not want to share their models or training data.

There are four primary components for on-premise deployments:

Back end: Deployed as a Docker container (exposes REST and web socket endpoints).

Front end: Deployed as a static website. It can be deployed on either an existing web server or as a Docker container provided as part of the installation package.

Database: The platform supports Microsoft SQL Server* and PostgreSQL*.

File storage: NAS or SAN connected to the backend for data files and optimized models.

In addition, a SaaS* portal exposes a limited version of the platform for companies that want to experiment with the technology before committing to an enterprise deployment of it.

DarwinAI and Intel® Optimizations for TensorFlow* Deliver up to 16.3X[1] Performance Increase on ResNet50 and up to 9.6X[1] on NASNet Workloads

Intel and DarwinAI engineers ran image classification performance tests with ResNet50 Convolutional Neural Network (CNN) and NASNet using the Intel® Optimization for TensorFlow* with Intel® Math Kernel Library (Intel® MKL) and Intel® MKL-DNN.

ResNet50 Inferencing Speedup 16.3X[1]

The ResNet50 network is an image classification-based CNN trained on over a million images from the ImageNet database (http://www.image-net.org). Using the Generative Synthesis platform with Intel technology and optimizations resulted in a 16.3X[1] improvement in inference performance over baseline measurements (images per second) for an Intel® Xeon® Platinum 8153 processor (Figure 1).

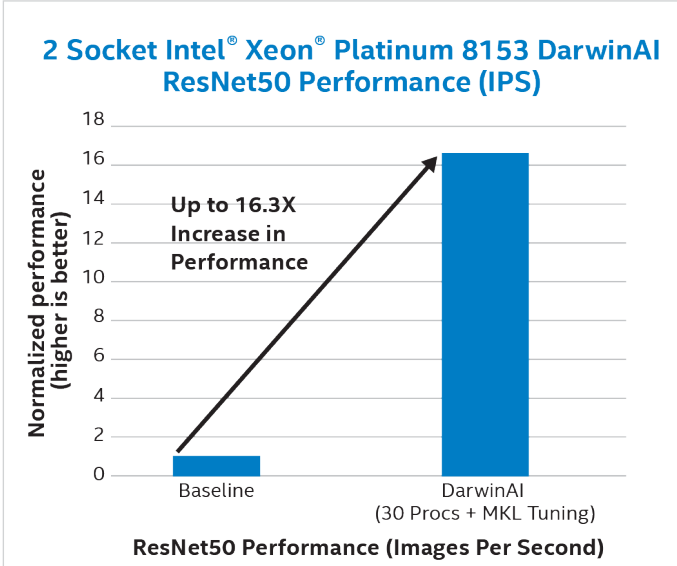

Figure 1. Improvements of ResNet50 performance with DarwinAI and Intel® Optimizations for TensorFlow*.

NASNet Inferencing Speedup up to 9.6X[1]

NASNet is another image classification-based CNN. Again, using the Generative Synthesis platform and the Intel® Optimization for TensorFlow with Intel MKL and Intel® Math Kernel Library for Deep Neural Networks (Intel® MKL-DNN) resulted in a 9.6X[1] improvement in inference performance over baseline (frames per second) for an Intel Xeon Platinum 8153 processor (Figure 2).

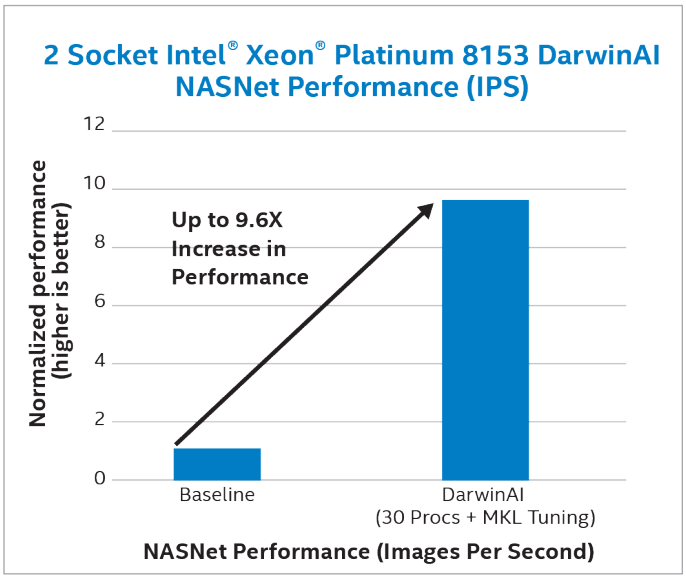

Figure 2. Improvements of NASNet performance with DarwinAI and Intel® Optimizations for TensorFlow*.

Conclusion

DNNs are complex systems that can be simplified using the DarwinAI Generative Synthesis platform. The platform leverages AI to optimize AI, resulting in smaller, high-performance networks that are also explainable. Generative Synthesis with Intel optimizations and technologies enabled up to a 16.3X inference performance improvement on the ResNet50 and up to 9.6X improvement on the NASNet image classification-based CNN models compared to their non-optimized baseline measurements.

The Intel optimizations leveraged the Intel MKL-DNN to take advantage of the technologies built into the Intel® Xeon® Scalable processor architecture.

The DarwinAI Generative Synthesis technology creates light, portable neural networks from existing model definitions with explainability. Benefits include the following:

- Faster inferencing can deliver insights quicker.
- Models may be less computationally expensive, enabling deployment to the furthest edge devices.
- Explainability helps developers better understand network decision-making, allowing them to target various system specifications and tune networks.
- Explainability can help smooth certification processes with regulating bodies.
- Significant speedup of inference times further enabled through Intel hardware and software.

For more information about the DarwinAI Generative Synthesis platform, visit https://darwinai.ca/

About DarwinAI

DarwinAI is a cutting-edge Artificial Intelligence startup based in Waterloo, Ontario, Canada. Their technology, stemming from years of scholarship by their academic team from the University of Waterloo, uses “AI to build AI.”

DarwinAI is a member of the **Intel® AI Builders Program**, an ecosystem of industry-leading independent software vendors (ISVs), system integrators (SIs), original equipment manufacturers (OEMs), and enterprise end users, which have a shared mission to accelerate the adoption of artificial intelligence across Intel® platforms.

1. Testing conducted comparing unoptimized with optimized DarwinAI* software using Intel® Xeon® Platinum 8153 processor. Testing done by Intel.

Configuration Details: DarwinAI EdgeSpeech, NASNet, ResNet50 - Throughput Performance on Intel® Xeon® Platinum 8153 Processor:

BASELINE: Tested by Intel as of 05/19/2019. 2 socket Intel® Xeon® Platinum 8153 Processor, 16 cores per socket, Ucode 0x200004d, HT On, Turbo On, OS Ubuntu 18.04.2 LTS, Kernel 4.15.0-46-generic, Total Memory 376 GB (12 slots/ 32GB/ 2666 MHz), BIOS SE5C620.86B.00.01.0015.110720180833, Deep Learning Framework: TensorFlow 1.13.1 (conda tensorflow-mkl), custom test data, tested using batches of 32 and 50, for NASNet and ResNet respectively, and n/a for Edgespeech, streams: 1 in all cases.

NEW: Tested by Intel as of 05/19/2019. 2 socket Intel® Xeon® Platinum 8153 Processor, 16 cores per socket, Ucode 0x200004d, HT On, Turbo On, OS Ubuntu 18.04.2 LTS, Kernel 4.15.0-46-generic, Total Memory 376 GB (12 slots/ 32GB/ 2666 MHz), BIOS SE5C620.86B.00.01.0015.110720180833, Deep Learning Framework: TensorFlow 1.13.1 (conda tensorflow-mkl), custom test data, tested using batches of 32 and 50, for NASNet and ResNet respectively, and n/a for Edgespeech (all optimized for latency), streams: 30 for all cases.

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.

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. 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: http://www.intel.com/performance.

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 the product when combined with other products. For more complete information, visit http://www.intel.com/benchmarks.

Performance results are based on testing as of May 2019 and may not reflect all publicly available security updates. See configuration disclosure for details.

Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Check with our system manufacturer or retailer or learn more at https://www.intel.com.

0619/BA/HBD/PDF ♻ Please Recycle