CASE STUDY

Intel® AI: In Production
AI-Enabling Technologies

DarwinAI Delivers Explainable AI Using OpenVINO™ Toolkit

Explainable AI maintains high performance while showing how neural networks build models using the Intel® Distribution of OpenVINO™ toolkit.

At a Glance

DarwinAI Generative Synthesis* platform brings transparency to show how neural networks work, and makes them run better.

- Understand how a neural network is reaching its decisions—important in both diagnostic and regulatory contexts.
- Accelerate deep learning development by collaborating with AI during the design process.
- Reduce computational requirements by generating neural networks optimized by AI (particularly useful in deploying deep learning at the edge).

"We use machine learning to probe and understand a neural network on a very fundamental level, and then build up a very sophisticated mathematical understanding of the network during that process. We then use AI a second time to generate an entirely new family of neural networks that is considerably more compact than the original, as good as the original from a functional standpoint, but works a lot faster. Specifically, by leveraging the Intel® Distribution of OpenVINO™ toolkit, we saw up to 6.18x improvement in FP32 data types and 10.51x in INT8 running on 2nd Gen Intel® Xeon® Scalable processor-based platforms."

- Sheldon Fernandez, CEO, DarwinAI

Today, Artificial Intelligence (AI) is democratized in everyday life, forecasted to grow from 12 billion U.S. dollars in 2017 to 52.2 billion U.S. dollars by 2021.[1] AI is shaping the next generation of digital business models and ecosystems, and is beginning to have a significant impact on society.

Meanwhile, the statistics portal Statista expects that revenues from the AI market worldwide will grow from 480 billion U.S. dollars in 2017 to 2.59 trillion U.S. dollars by 2021.[2] Gartner identifies AI as an inescapable technology among the Gartner Top 10 Strategic Technology Trends for 2018. Along with immersive experiences, digital twins, event-thinking and continuous adaptive security, AI solutions are shaping the next generation of digital business models and ecosystems[3], and the proliferation of AI is having a significant impact on society. AI has already become ubiquitous and we have become accustomed to AI making decisions for us in our daily life, from product and movie recommendations to friend suggestions on social networks and tailored advertisements on search result pages. However, in life-changing decisions, such as disease diagnosis or autonomous vehicles, it is important to know the reasons behind such a critical decision, and the crucial need for explaining AI-generated outcomes becomes fully apparent.

Challenge

Though they appear powerful in terms of results and predictions, AI algorithms—especially deep neural networks—suffer from a series of drawbacks. Specifically, deep neural networks are:

- Difficult to build, requiring expertise to construct
- Difficult to run, necessitating significant computational resources
- Difficult to explain, with a lack of insight into how they reach their decisions

The opacity of neural networks is a byproduct of its tremendous complexity. This limitation is highly problematic, since if one doesn't know how a model is reaching a decision, one can't predict how and when it will fail. This is especially troubling in critical, life-and-death scenarios, as entrusting important decisions to a system that cannot explain itself is both precarious and dangerous.

Solution

To address this challenge, Explainable Artificial Intelligence (XAI) aims to make AI transparent using techniques that produce more explainable models while maintaining high performance levels.

DarwinAI Generative Synthesis* technology pioneers the use of AI to build AI. Employing traditional forms of Machine Learning, the platform observes a neural network and then uses those observations to build new, highly-optimized versions

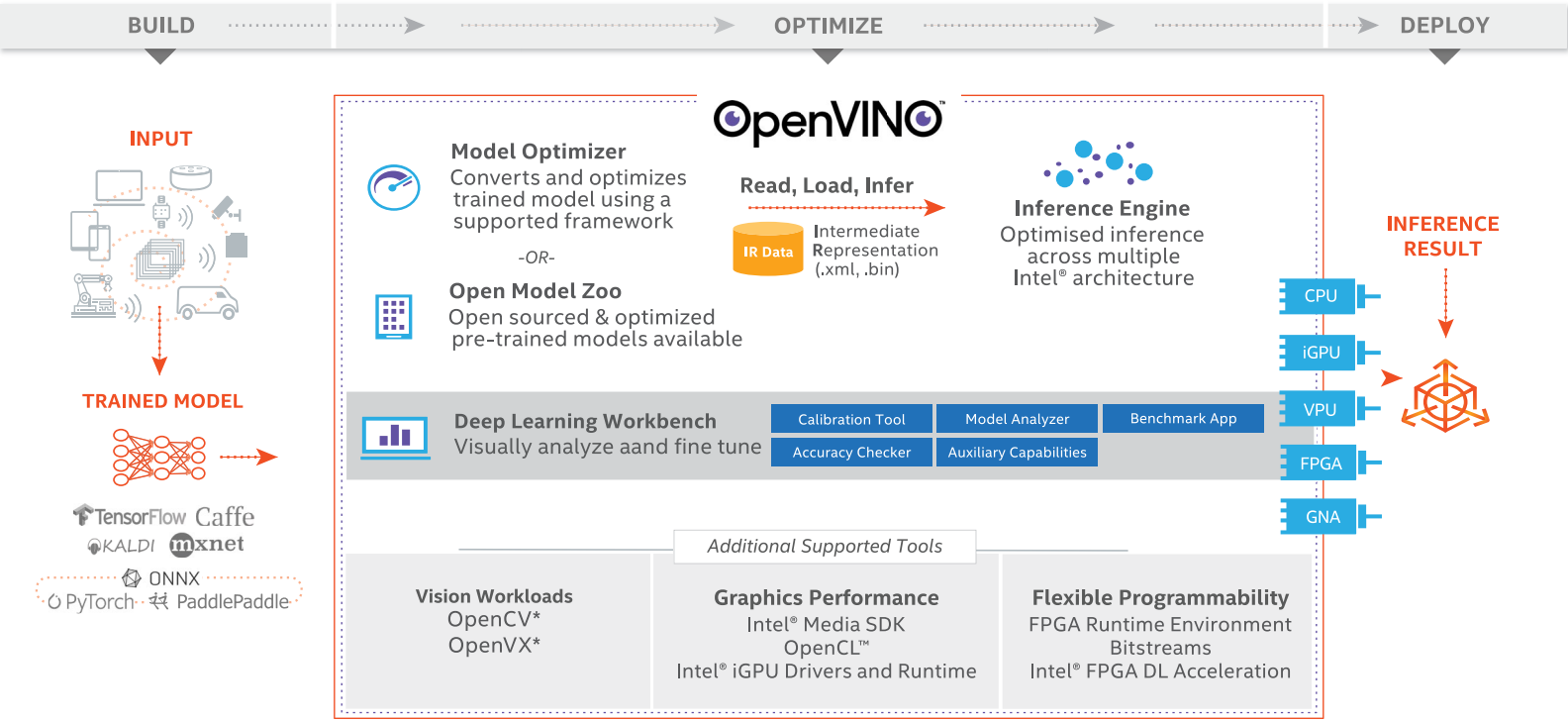
of that network. This reduces the complexity of designing high-performance deep learning solutions, generates highly optimized models for edge computing, and also facilitates XAI to understand why a network makes the decisions it does.

In addition, Generative Synthesis* is complementary to the Intel® Distribution of OpenVINO™ toolkit and low-level software components of the toolkit, such as Intel® Optimizations for TensorFlow*, and Deep Neural Network Library (formerly Intel® Math Kernel Library for Deep Neural Networks or Intel® MKL-DNN)—making it optimized for deploying deep learning applications to Intel® architecture.

In fact, DarwinAI is utilizing the Intel Distribution of OpenVINO toolkit on an image classification workload with Resnet50 Convolutional Neural Network (CNN) comparing the use of an original TensorFlow model on both FP32 and INT8 data types.

Figure 1 | Under the Hood of Intel Distribution of OpenVINO toolkit

The Intel Distribution of OpenVINO toolkit is a free software kit that helps developers and data scientists speed up computer vision workloads and streamline deep learning deployments from the network edge to the cloud.



Result

Utilizing the Intel Distribution of OpenVINO toolkit on an image classification workload with a Resnet50 CNN, Intel and DarwinAI compared the use of an original TensorFlow model on both FP32 and int8 data types with optimized networks produced by DarwinAI. A comparison of inference performance on 2nd Gen Intel® Xeon® scalable processor-based platforms found up to 6.18x improvement in throughput using the Intel Distribution of OpenVINO toolkit with FP32 and up to 10.51x improvement using Intel Distribution of OpenVINO toolkit with INT8 compared to just using TensorFlow* 1.13.1.

See Figure 2.

Compare Inference Performance on 2nd Gen Intel® Xeon® Scalable Processor

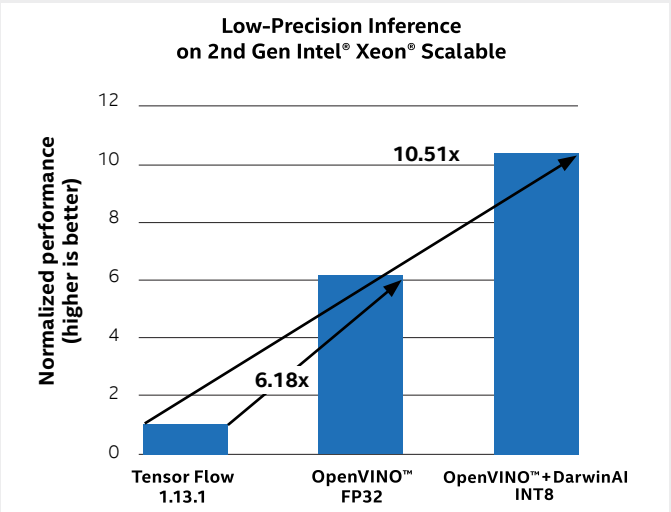


Figure 2 | ResNet50 Performance (Image Per Second)

In addition, networks built using the Generative Synthesis* platform coupled with Intel® Optimizations for TensorFlow* were able to deliver up to 16.3X performance increase on ResNet50 and up to 9.6X on NASNet workloads, improving inference performance over baseline measurements (images per second) for an Intel® Xeon® Platinum 8153 processor.

By utilizing the Generative Synthesis* platform, an automotive end user accelerates development and reduces their time to market while saving money on the costs associated with validating their neural network. The platform also helped reduce the size and improved performance of the perception network behind their autonomous vehicle systems.

See Figure 3.

2 Socket Intel® Xeon® Platinum 8153
DarwinAI ResNet50 Performance (IPS)

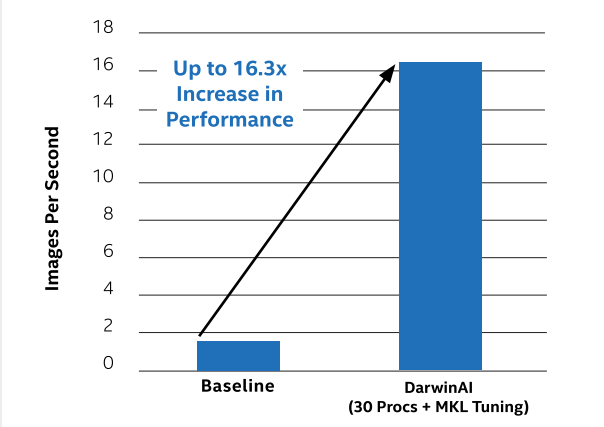


Figure 3 | ResNet50 Performance (Image Per Second)

Key Benefits

- **Accelerated development and reduced time to market:** The platform reduced the number of talent hours and GPU hours required for system development (200 talent hours to 17 hours, and 10,400 GPU hours to 760 hours).
- **Faster training and reduced validation costs:** There was an estimated 610,000€ in savings for test and validation tasks on a five-GPU cluster instance.
- **Optimization of neural networks optimization:** The platform reduced the size and increased performance for neural networks at the center of their autonomous systems.

Unparalleled Optimization and Explainability

In addition to Explainability, the Generative Synthesis* platform provides unparalleled levels of optimization. While common approaches to neural network optimization – pruning and precision weight reduction – reduce model size, they usually do so at the expense of accuracy. Generative Synthesis*, by contrast, takes apart the prevailing model and reconstructs more compact, efficient and explainable variants of said model that maintain, and in some cases improve, accuracy while reducing inference time.

When used in combination, the design, optimization and XAI elements of Generative Synthesis* enable human-machine collaboration to improve model performance and robustness. Often, the platform can uncover problems that wouldn't occur with a human developer.

For example, in the case of an automotive end user, DarwinAI helped diagnose a problem with an autonomous vehicle system that exhibited bizarre behavior where the car would turn left with increasing regularity when the sky was a certain shade of purple. Generative Synthesis* helped determine that the training for certain turning scenarios had been conducted in the Nevada desert when the sky was a particular hue. Unbeknownst to its human designers, the neural network had established a 'nonsensical correlation' between its turning behavior and the celestial tint.

Conclusion

The AI building AI technology delivered by DarwinAI assists with key challenges when building deep neural networks. By illuminating how networks make their decisions, Generative Synthesis* helps developers implement more robust models. Moreover, the high levels of optimization facilitated by the platform constitute an essential tool in deploying AI at the edge. Used in combination, these offerings are a powerful complement to Intel® architecture paired with and powered by Intel® Distribution of OpenVINO™ toolkit.

Solution Ingredients

» Intel® Distribution of OpenVINO™ toolkit

» Deep Neural Network Library (formerly Intel® Math Kernel Library-DNN or MKL-DNN)

» Intel® Optimizations for Tensorflow*

Learn More

» DarwinAI

» Build Explainable AI with DarwinAI: Podcast

» Generative Synthesis* Platform: Solution Brief

About Intel® AI: In Production

Intel® AI: In Production is an ecosystem focused on reducing deployment complexities, promoting partner AI offerings, and increasing collaboration between its partners.

DarwinAI Generative Synthesis* platform is an AI-based platform that provides deep learning neural networking solutions for enterprise applications.

Combined with the Intel® AI: In Production ecosystem, partners can accelerate the development of offerings powered by the Intel® Distribution of OpenVINO™ toolkit

About DarwinAI

DarwinAI is an alumni graduate of the 2018 Creative Destruction Lab (CDL) program, was recently included on insideBIGDATA's list of top 50 impactful companies for Q4 2019, and was also named a 'cool vendor' by Gartner in their Enterprise AI Governance and Ethical Response Report.

The company's technology—the byproduct of years of academic study from the University of Waterloo—has been used in numerous industrial contexts including autonomous vehicles, consumer electronics, aerospace and military, security, and financial services. For more information, visit darwinai.ca.

Find the solution that is right for your organization
Contact your Intel representative or visit:
www.intel.com/ai-in-production

DARWIN AI

intel®