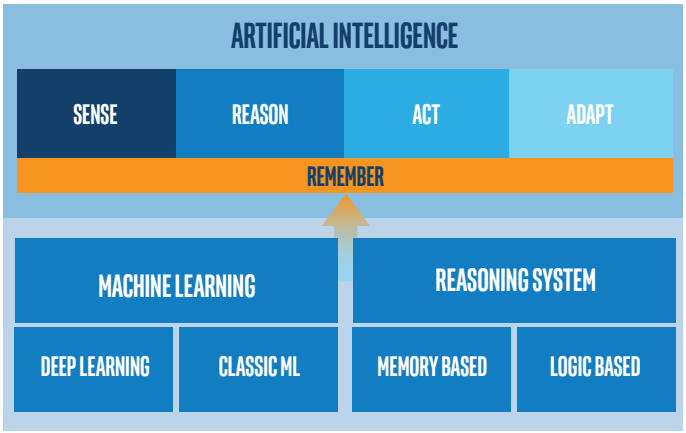
PRODUCT BRIEF

Cloud Service Providers, Surveillance & Security, Retail, Manufacturing

Intel® Deep Learning Inference Accelerator

Accelerate CNN Workloads with Turnkey Inference Solution



Intel Deep Learning Inference Accelerator

Hardware

PCIe add-in card powered by Intel® Arria 10 FPGA

Software

Integrated deep learning stack with industry-standard libraries and frameworks

Intellectual Property

Optimized CNN algorithms supporting multiple network topologies

Artificial Intelligence: The Next Wave of Computing

In our smart and connected world, machines are increasingly learning to sense, reason, act, and adapt in the real world. This is artificial intelligence (AI). Machine learning, deep learning and reasoning-based systems are leading approaches to AI.

Machine Learning

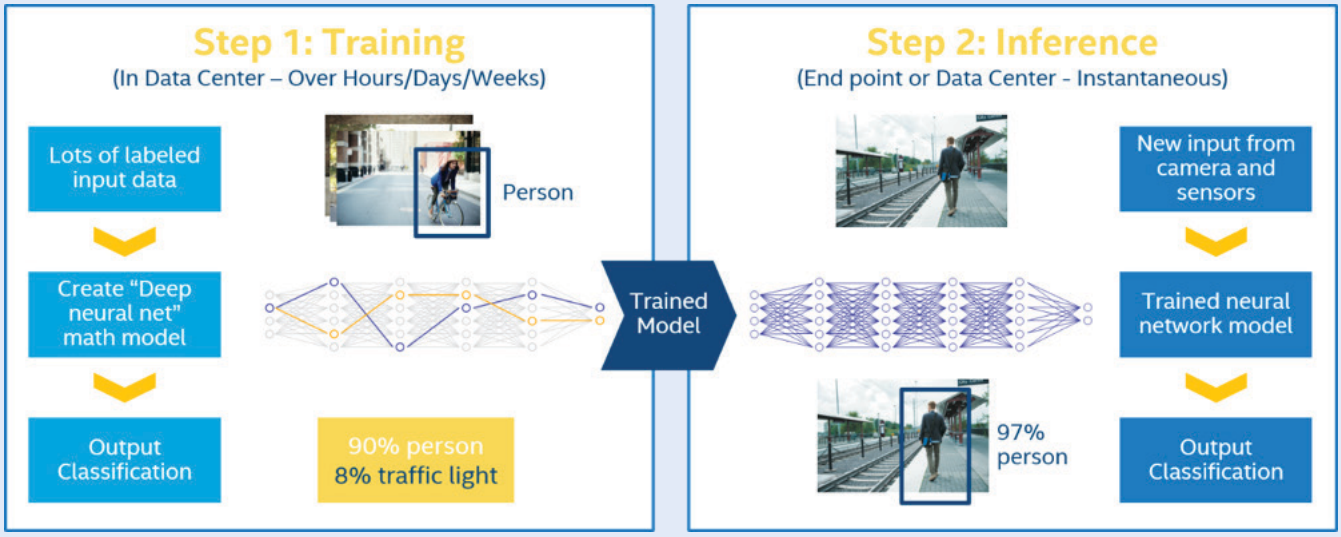
Machine learning is the dominant approach to AI today. Machine learning involves using a variety of algorithms that "learn" from analyzing data and improve performance based on real-world experience. Most machine learning techniques are considered "conventional," because many have been in use for well over a decade, and they generally rely on mathematical or statistical algorithms to perform regression, decision trees, classification, clustering, and many more functions. This form of learning is often optimal when the dataset is tabular, for example, trying to predict the likelihood of a sales deal closing based on historical activities and interactions with the prospect, or recommending a new item to buy based on previous purchases.

Deep Learning

Deep learning is a rapidly emerging branch of machine learning, which relies on large data sets to iteratively "train" many-layered deep neural networks (DNN) inspired by the human brain. Trained neural networks are used to "infer" the meaning of new data, with increased speed and accuracy for processes like image search, speech recognition, natural language processing, and other complex tasks. Deep learning is used for the facial recognition and tagging features on social media, voice recognition on our smartphones, autonomous driving, and personal assistants. Convolutional Neural Networks (CNN) are used for image classification, detection and computer vision applications.

Intel® Deep Learning Inference Accelerator (Intel® DLIA)

The Intel® DLIA is a turnkey inference solution that accelerates CNN workloads for image recognition, classification and computer vision applications. It integrates industry-leading software frameworks and libraries, such as the Intel® Math Kernel Library for Deep Neural Networks (Intel® MKL-DNN) and Intel® Caffe*, to enable simple software-level programmability, while the PCIe*-based accelerator card increases throughput and power efficiency.



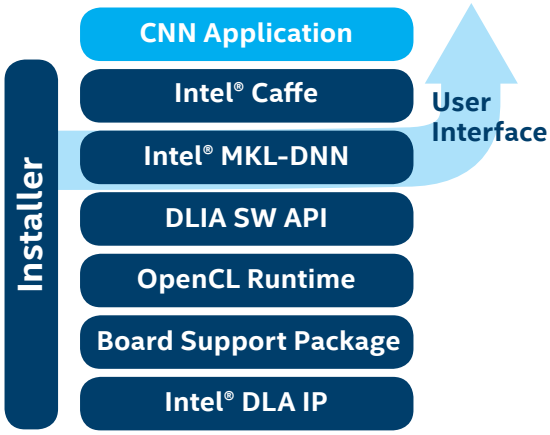
Step 1: Training

Today training runs in the data center and the key metric time to train is in hours/days/weeks. Labeled data is fed into a blank model. With each piece of data, the model adjusts and updates its weights, making it incrementally more accurate. This can continue until an acceptable level of accuracy is reached. In the case of image recognition, this requires a large amount of data – typically millions of images. Often, this step is repeated with a number of model architectures, or network topologies, to find out which model topology performs the best.

Once a model has been chosen, and enough data has been processed to create stable and accurate model weights with acceptable accuracy, it is then deployed into the second phase: inference.

Step 2: Inference

Inference (sometimes called scoring or classification) happens instantaneously at the edge or in the data center, such as when a new photo is uploaded for inspection. In this step, a single, unsupervised data point – an image where the subject is unknown, for example, is fed through the completed model and characterized. The output is a prediction of what is contained in the image. For inference throughput, performance efficiency and total cost of ownership (TCO) are critical metrics.



Software Ecosystem

Intel DLIA includes an optimized software ecosystem for CNN algorithms. The FPGA is preprogrammed with Intel Deep Learning Accelerator IP (DLA IP) to accelerate CNN primitives. These primitives are enabled through Intel MKL-DNN which provides a unified deep learning API and is optimized for performance throughout the stack. Applications can be built through the Intel DLIA software stack using the Caffe framework or MKL-DNN primitive API.

In the rapidly evolving space of AI, algorithms, frameworks and topologies are changing at breakneck speed. FPGAs, or field programmable gate arrays, can be reprogrammed even once deployed. Intel DLIA offers optional software and IP upgrades, such as: new primitives, models, lower precision computation, frameworks and features, on the same hardware. To provide the best user experience and simplify ease of use, Intel DLIA is only functional with the included IP packages and is not reprogrammable to custom IP.

Benefits

Intel DLIA is a packaged end-to-end offload acceleration solution for deep learning deployment that reduces the overhead of complex hardware and software integration. FPGA-based accelerators offer scalable throughput gains, at power efficiencies several times better than a CPU alone – lowering total cost of ownership for high throughput systems.

DLIA's software ecosystem allows customers and end users to create cutting edge, value added applications instead of developing non-core IP or custom software interfaces. Intel MKL-DNN and frameworks, such as Intel Caffe, offer a unified software framework that works across Intel® architectures, enabling code reuse and portability across the portfolio of Intel-based offerings. Intel DLA IP enables flexible acceleration to a variety of CNN-based topologies, including: AlexNet, GoogLeNet, VGG-16, LeNet, CaffeNet and SqueezeNet, all reconfigurable through software.

Intel® DLIA Use Cases

Cloud Service Providers

Filter Inappropriate Content
Track Product Photos

Surveillance and Security

Facial Recognition
License Plate Detection

Manufacturing

Detect Defects
Fault-tolerant mfg Lines

Retail

Track Store Traffic
Monitor and Track Visual Inventory

Specifications

Intel® Deep Learning Inference Accelerator	Details
Hardware	
Form Factor	Full-length, full-height, single wide PCIe* card
FPGA	Intel® Arria 10 @ 275MHz
TFLOPS	Up to 1.5
Memory	2 Banks 4GB x64 DDR4
PCIe Express* Configuration	Gen3 x16 host interface; x8 electrical, x16 power & mechanical
TDP	50-75W
Cooling	Active
Operating temperature	0-85C
Operating System Support	CentOS* 7.2
Cards supported per host/system	2/4
Software	
Network topologies	AlexNet, GoogleNet, CaffeNet, LeNet, VGG-16, SqueezeNet, custom-developed
Framework	Intel® Caffe
Pre-programmed IP	Intel® Deep Learning Accelerator IP (DLA IP) Accelerates CNN primitives in FPGA: convolution, fully connected, ReLU, normalization, pooling, concat. Networks beyond these primitives are computed with hybrid CPU+FPGA
Libraries	Intel® Math Kernel Library for Deep Neural Networks (MKL-DNN)
Upgrades	Optional PCIe-based software updates to lower precision compute, new primitives and enabling new features

Order Code

Product	Order Code	Type	Description
Intel® DLIA	DLP3A115P3XX	PCIe Add-in Card	End-to-end offload acceleration solution for deep learning deployment powered by Intel® Arria 10 FPGA

Product does not include memory, processors, or hard drives. For compatibility information please refer to the configuration guide at **www.intel.com/support**.

Results have been estimated or simulated using internal Intel analysis or architecture simulation or modeling, and provided to you for informational purposes. Any differences in your system hardware, software or configuration may affect your actual performance.

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 your system manufacturer or retailer or learn more at **ark.intel.com**.

 0417/JL/PDF 335776-001 US