INDUSTRIAL COMPUTER PRODUCTS

...by people who care!

For IIoT Developers

AI TRAINING & INFERENCE SYSTEMS

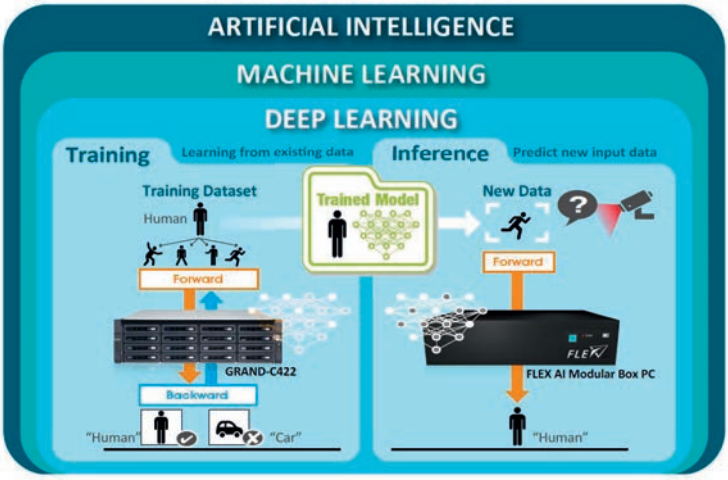
+ Application deployment at the edge
Hardware and software compatibility distributed across all AI systems to realize end-to-end intelligence with widely adopted deep learning frameworks.

+ Flexible
Highly interconnected with sensors and peripherals and flexibly expandable due various standardized interfaces and bus systems.

+ High-performing and scalable
Intel® processor and acceleration technologies.
Massive system memory and storage capacity.

How does deep learning work?

Deep learning (DL) is a machine learning technique that is based on deep neural networks and recurrent neural networks architectures. These artificial neural networks are comparable with the structure of the human brain. The underlying learning procedure is realized by using representations of features directly from data such as images, text and sound. These representations are the result of abstracting input data in multiple layers and on different levels, which form a concise network. DL is applied for instance in fields of computer and machine vision, speech and audio recognition or social network filtering.



In some cases the performance of deep learning algorithms can be even more accurate than human judgement.

There are **three steps** to successfully conduct deep learning projects:

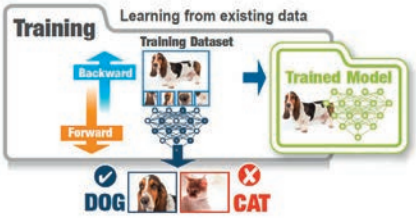
1. Data acquisition

2. Training

3. Inference

Data acquisition & training

Before the training of machine vision applications begins, **data acquisition** is required. A huge and varied set of data from any kind of source like web, cloud, sensors and images must be available. Then **data aggregation** and labeling takes place in order to be able to classify e.g. animal's images accurately. Based on the aggregated data the DL model gets developed and trained. The **training** is being done by a learning procedure that consists of layers. The first input layer contains raw data e.g. a matrix of pixels which becomes automatically more abstract in the next layers (hidden). Step by step specific features (e.g. edges, eyes, nose) are encoded until the complete animal is recognized in the final output layer. The result is a dog can be distinguished from a cat.



GRAND-C422 AI Training System

The AI training system **Grand-C422** is dedicated for these tasks because it offers a wide range of slots for storage expansion, acceleration cards and video capture, Thunderbolt™ or PoE add-on cards for unlimited data acquisition possibilities. In order to develop a useful training model, existing and widely used deep learning training frameworks such as Caffe, TensorFlow or Apache MXNet are recommended. These facilitate the definition of the apt architecture and algorithms for a distinct AI application.

Supported Software

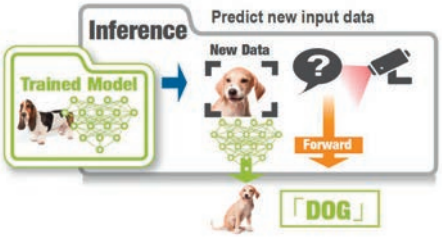
TensorFlow

Caffe

mxnet

Inference and optimization

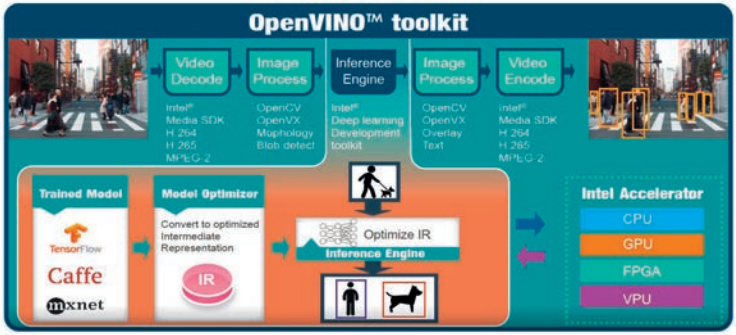
Is the DL model ready for trial, it can be transferred to the inference system TANK-870AI for performance optimization and the execution of inference tasks. Inference is known as reasoning in steps, whereby existing knowledge is used to draw logic conclusions. In the field of AI, neural networks undertake this “prediction” and “scoring” by passing new data through a trained model to compute results for each query.



TANK-AI Inference System

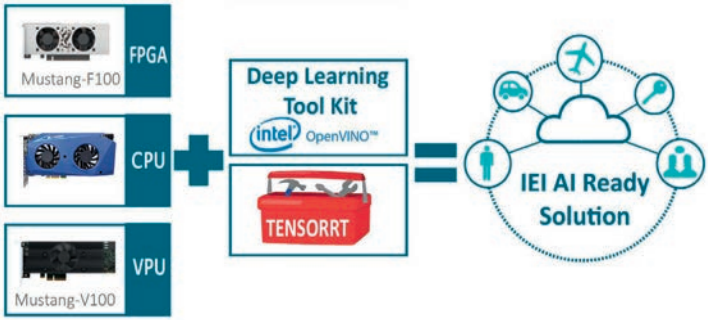
One major advantage of the TANK-870AI is the preinstalled open-source developer toolkit Open Visual Inference Neural Network Optimization (OpenVINO™). It is compatible with widely adopted DL frameworks, optimizes the DL model via an integrated model optimizer and inference engine (runtime) and accelerates the deployment process of the inference solution to the edge by pre-trained models, samples, and demos.

Included Software



Computing accelerators

In addition, the performance of optimized inference models can be further enhanced by adding heterogeneous low profile computing acceleration cards such as the Intel® field programming gate arrays (FPGA) or the Intel® Movidius® vision processing units (VPU). An alternative to the openVINO™ toolkit is TensorRT. The combination of GRAND-C422, TANK-870AI, the accelerator cards and a DL toolkit form IEI's AI ready solution.



IEI AI Ready Solution

ICP offers three different acceleration cards. Whereby the Mustang-V100-MX8 is based on Intel® Movidius Myriad X and the Mustang-F100-A10 is based on Intel® Arria 10GX 1150 FPGA. Both are designated for inference enhancement. The CPU acceleration card Mustang-200 combined two Intel® Core ULT CPUs and offers additional inference performance.

Intel Vision Accelerator Design Products

Available AI systems

Data acquisition systems	
Applications	Data acquisition of sensor data for training systems
Processing Unit	CPU / from Intel® Bay Trail to Intel® Kaby Lake
RAM	up to 64GB

Training systems	
Applications	Processing of data to create training models
Processing Unit	CPU / Intel® Xeon® W
RAM	up to 256GB

Inference systems	
Applications	Machine Vision, Object detection,
Processing Unit	CPU / Intel® Coffee Lake, Intel® Kaby Lake
RAM	up to 64GB

CPU computing accelerator	
Applications	Training System Optimizer
Processing Unit	CPU / 2x Intel® Kaby Lake
RAM	32GB
Size	full-height full-length double-width PCIe

FPGA computing accelerator	
Applications	Inference System Optimizer
Processing Unit	FPGA / Intel® Arria® 10
RAM	8GB
Size	half-height half-length double-width PCIe

VPU computing accelerator	
Applications	Inference System Optimizer
Processing Unit	VPU / 8x Intel® Myriad™ X
RAM	-
Size	half-height half-length single-width PCIe

ICP Deutschland GmbH
Mahdenstr. 3 | D-72768 Reutlingen | Tel: +49 (0) 7121 14323-20
sales@icp-deutschland.de | www.icp-deutschland.de