SOLUTION BRIEF

Integrated AI Training and Inference for the Edge

INTRODUCTION

Training and inference demand massive compute resources that utilize expensive and power-hungry GPUs. Consequently, deep learning is performed in the cloud or in large on-prem data centers. Training new models take days and weeks to complete, and inference queries suffer from long latencies of the round-trip delays to and from the cloud.

Yet, the data which feeds into the cloud systems, for updating training models and for inference queries, is generated mostly at the edge – in stores, factories, terminals, office buildings, hospitals, city facilities, 5G cell sites, vehicles, farms, homes and hand-held mobile devices. Transporting the rapidly growing data to and from the cloud or data center leads to unsustainable network bandwidth, high cost and slow responsiveness, as well as compromises data privacy and security and reduces device autonomy and application reliability.

Deep-AI has developed a unique, integrated, and efficient training and inference deep learning solution for the edge. With Deep-AI, application developers can deploy an integrated training-inference solution with real-time retraining of their model, in parallel to online inference on the same device.

PRODUCT OVERVIEW

Deep-AI's s/w solution runs on off-the-shelf FPGA cards, eliminating the need for GPUs, and provides a **10X gain** in performance/power or performance/cost versus a GPU. The FPGA hardware is completely under-the-hood and transparent to the data scientists and developers designing their AI applications. Standard deep learning frameworks are supported including Tensorflow, PyTorch and Keras.

Core Technology

- Quantized Training at 8-bit fixed point
- Training at high sparsity ratios

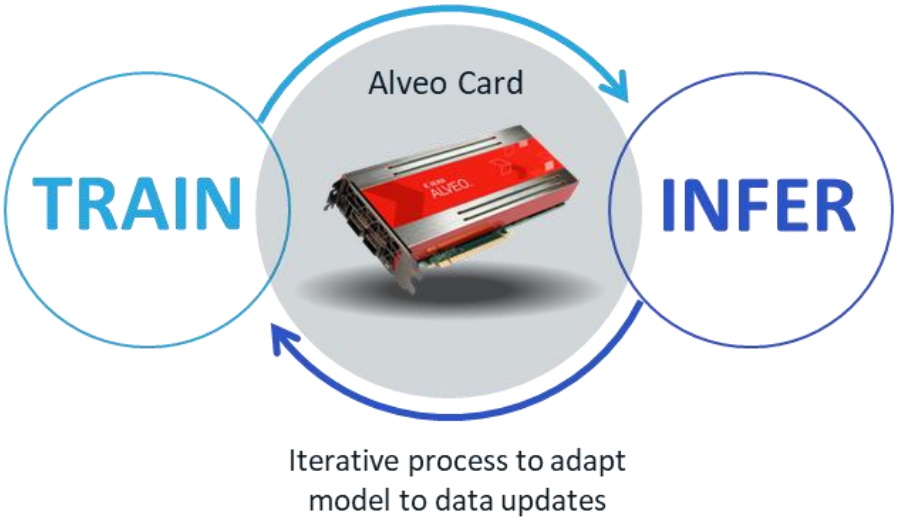
Benefits

- Same h/w for training & inference
- Faster training
- 10X gain in performance/power or performance/cost vs. GPU
- Training output is inference ready
- Hands-free, seamless switch between training and inference
- Scalable, Secure & Reliable: no need to send data back to cloud or data center

Adaptable. Intelligent.

SOLUTION OVERVIEW

Deep-AI's solution runs on Xilinx Alveo PCIe cards, certified and available on a variety of standard servers from leading server vendors. The **same hardware** is used for inference and retraining of the deep learning model, allowing an on-going iterative process that keeps the model updated to the new data that is continuously generated.



Iterative process to adapt model to data updates

- **Fixed-point 8-bit output feeds directly to inference**
- **No manual post-training processing / calibration**
- **No loss of accuracy from training to inference**

Supported Neural Networks:

- CNNs for imaging applications including all popular neural layers, convolutions, max/average pooling, residual shortcut, batch norm
- Resnet and Mobilenet for Classification
- Yolo, TinyYolo and SSD for Object Detection
- MLP
- More to come

TAKE THE NEXT STEP

For more information or to request a demo please visit us at:
https://deep-aitech.com/
or send an email to: info@deep-aitech.com

XILINX