Deep-AI Launches Industry-First Integrated AI Training and Inference Solution for the Edge

Featuring breakthrough technology for training at 8-bit fixed-point coupled with high sparsity ratios to enable deep learning at a fraction of the cost and power of GPU systems for fast, secure, and scalable AI deployments at the edge

Oct 7, 2020. Caesarea, Israel

Deep-AI Technologies is emerging from stealth and launching the industry's first integrated training and inference solution for deep learning at the edge. With Deep-AI, every inference node at the edge also becomes a training node, enabling faster, cheaper, scalable, and more secure AI versus today's cloud-centric AI approach.

Deep-AI's 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 the developers designing their AI applications. Standard deep learning frameworks are supported including Tensorflow, PyTorch and Keras.

Training deep learning models and servicing inference queries demand massive compute resources delivered by expensive, power-hungry GPUs, and consequently deep learning is performed in the cloud or in large on-premise data centers. Training new models takes 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 the training models and the 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.

To overcome these limitations, Deep-AI has uniquely developed an integrated, holistic, 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.

At the core of Deep-AI's technology is the ability to train at 8-bit fixed-point coupled with high sparsity ratios at training time, as opposed to 32-bit floating-point and no sparsity which is the norm today with GPUs. These two technological breakthroughs enable AI platforms that are superior in performance, power, and cost. When realized into an ASIC they can drive a 100X efficiency in silicon area and power over GPUs.

Innovative algorithms compensate for the lower precision of 8-bit fixed-point and the high sparsity and minimize any reduction in training accuracy. For edge applications, where the use cases typically call for the retraining of pre-trained models with incremental data updates, the training accuracy is fully maintained in most cases and with minimal reduction in other cases.

Furthermore, in most systems today training is done at 32-bit floating-point while there is a growing desire to run inference at 8bit fixed-point. In these cases, one needs to manually run challenging as well as time and resource consuming quantization processes to convert the 32-bit training output into an 8-bit inference input. Moreover, this conversion often results in loss of accuracy. Because Deep-AI's training is done in 8-bit fixed-point it is inference-ready by design and feeds directly to

inference. No manual intervention nor processing is needed to quantize the training output before inference and no loss of accuracy is experienced from moving from training to inference.

Deep-AI's solution uses FPGAs, which are rapidly growing in adoption for a wide variety of acceleration workloads. Recent advancements in deep learning enable inference with 8-bit fixed-point number formats and enable very low-latency inference on FPGAs. Deep-AI's breakthrough technology takes a huge step forward by also enabling training on FPGAs with 8-bit fixed-point number formats and running both training and inference on the same FPGA platform.

Deep-AI's solution is available today for on-premise deployments on standard off-the-shelf FPGA cards from Xilinx and leading server vendors. The solution will also be available on Xilinx cloud-based FPGA-as-a-service instances in the first quarter of 2021.

Collaboration with Xilinx, Dell Technologies and One Convergence

Deep-AI's solution runs on Xilinx Alveo accelerator cards, PCIe add-in 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.

"Deep-AI has demonstrated impressive capability to address the challenges of fixed-point training for deep learning models" said Ramine Roane, Vice President of Software & AI Solutions Marketing at Xilinx. "Xilinx is excited to be working with Deep-AI to bring to market a training solution based on our adaptive platforms."

Deep-AI working with Dell Technologies has validated the PowerEdge R740xd rack servers with pre-installed Xilinx Alveo cards and sample network models and data sets, in particularly for the retail and manufacturing markets.

In addition, Deep-AI has partnered with One Convergence, to offer customers the Deep-AI solution integrated with the One Convergence DKube complete end-to-end enterprise MLOps platform.

"We are happy to be partnering with Deep AI and offer cost effective, integrated training and inference acceleration solutions to our customers though our DKube platform" said Ajai Tyagi, Senior Director, Marketing and Sales. " DKube (https://www.dkube.io) is a modern Kubernetes-based platform based on open standards such as Kubeflow and MLFlow, and it addresses the critical needs of the AI communities for a common, integrated MLOps workflow, especially those that want to deploy on-prem and/or hybrid models."

About Deep-AI Technologies

Deep-AI Technologies delivers accelerated and integrated deep-learning training and inference at the network edge for fast, secure, and efficient AI deployments. Our solutions feature breakthrough technology for training at 8-bit fixed-point coupled with high sparsity ratios, to enable deep-learning at a fraction of the cost and power of GPU systems. For more information and scheduling a demo visit https://deep-aitech.com