The Linley Group

MICROPROCESSOR report

Insightful Analysis of Processor Technology

DEEP AI RETRAINS NETWORKS AT THE EDGE

Startup Uses FPGA to Bring DNN Retraining to the Infrastructure Edge

By Bob Wheeler (November 2, 2020)

Startup Deep AI is enabling infrastructure-edge hardware to perform both retraining and inference. Otherwise, before running AI inference, customers must retrain their neural network using costly on-premises hardware or in the cloud. Retraining refers to adapting a pretrained model to customer data sets rather than training from scratch. Deep AI initially targets applications in markets such as retail, manufacturing, smart cities, and health care.

The company licenses its software and RTL for use with FPGA-based accelerator cards from Xilinx. Its initial offering is for on-premises deployment in Alveo U50 cards and is available now. In 1Q21, it plans to enable FPGA-as-a-service instances based on the Alveo U250 to use its software/RTL.

Delivering acceptable training time from such modest hardware requires a secret sauce, which Deep AI provides through reduced-precision math and sparsity. Inferencing with 8-bit integer (INT8) data is already popular, but it requires quantizing models trained at higher precision, such as 32-bit floating point (FP32). In Deep AI's system, inferencing directly employs the trained model without any intermediate conversions. The startup claims these models, including sparsity, are up to 95% smaller than the original model trained using FP32 without sparsity.

By retraining at the edge, customers needn't send data to the cloud, improving privacy and security. Although researchers have studied INT8-based training, frameworks that achieve acceptable accuracy on a range of neural networks have yet to emerge. If Deep AI can deliver on its claims, it will set itself apart from other vendors targeting the infrastructure edge.

Retraining Is Different

Based in Israel, Deep AI was founded in 2017 by veterans of network-processor vendor EZchip, which Mellanox (now Nvidia) acquired in 2016 (see *MPR 10/19/15,* "EZchip Gives Mellanox Brains"). Cofounders Amir Eyal and Moshe Mishali serve as CEO and CTO, respectively. Eyal's previous role was VP of Business Development and Marketing at EZchip, where Mishali was an algorithm architect. Deep AI has raised $8 million from venture firms S Capital and i3 Equity as well as strategic investor Xilinx. Its 19 employees develop algorithms, RTL, and software.

Intel, Xilinx, and other FPGA vendors have focused on AI inference because low-precision integer math is a good fit for their chips. The two leaders supply FPGA-based accelerator cards as well as associated tools and intellectual property for AI inference. The initial Alveo line included the U280, which sports 8GB of HBM2 but dissipates up to 225W and requires two PCIe slots (see *MPR 2/18/19,* "Xilinx Delivers Server Acceleration"). Last year, Xilinx added the slimmed down U50, which omits DDR4 DRAM and provides less programmable logic than the U280 but fits in one slot and consumes 75W (maximum). The

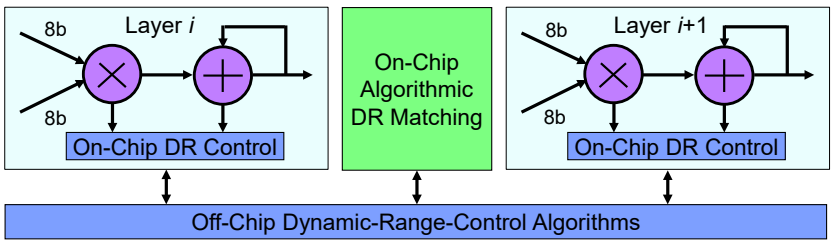

Figure 1. Training using 8-bit fixed-point data. DR=dynamic range. Deep AI employs proprietary matching algorithms to preserve dynamic range during back propagation, which only serves in training.

Price and Availability

Deep AI offers annual licenses at $500 per accelerator. Software for the Xilinx Alveo U50 is available now. The company plans to license RTL and software for use with FPGA cloud instances beginning in 1Q21. For more information, access *www.deep-aitech.com*.

U50 delivers two-thirds the INT8 performance of the U280, achieving about 16 TOPS.

Easing server integration, the U50 was Deep AI's obvious choice for its first hardware target. As Figure 1 shows, the startup performs retraining through the same INT8 multiply-accumulate functions that inference employs, and it implements proprietary dynamic-range control at each layer. It takes advantage of the host CPU to configure the matched filter that sits between layers. During both forward and back propagation, the FPGA data path operates without host-CPU involvement. When handling inference, the same data path operates without adjusting the interlayer dynamic-range matching.

The critical factor enabling INT8 training is that the company primarily intends it for retraining a pretrained model. For example, customers can take the ResNet-50 model pretrained on the ImageNet data set and retrain it using their own data set. Note that models aren't retrainable with only incremental data—a process called *incremental learning*—as they may forget prior learning. Typically, customers need a subset of the original data set (e.g., ImageNet) plus their own proprietary data set to retrain the model.

Training a model from scratch requires a huge dynamic range because the weights are unknown. Employing a stable pretrained model results in a bounded problem that's easier to handle at reduced precision. In fact, Deep AI claims to deliver 0–1% accuracy degradation compared with FP32 training for the models it supports.

Although many inference accelerators now implement some form of sparsity, Deep AI uniquely applies unstructured sparsity during training. Enabling sparsity usually degrades model accuracy, so the company allows customers to tune speed-up versus accuracy to meet their needs. At the extreme, it claims to achieve 90% sparsity for the ResNet-50 model trained with ImageNet. That is, the trained model needs only 10% as many weights as an unpruned model, while accuracy degrades 1% or less. Deep AI's 95% model-size-reduction claim requires 80% sparsity (combined with INT8 instead of FP32 data). The startup also employs AI-specific algorithms to compress activations, maximizing memory bandwidth for streaming data.

Abstraction Without Synthesis

It's one thing to develop unique technology, but it's another to deliver the technology in a form that unsophisticated customers can adopt. The enterprise customers Deep AI targets generally select from a relatively small set of models developed for their application, and they lack the resources to customize a model. To serve these customers, the startup is focusing on convolutional neural networks (CNNs) and simpler multilayer perceptrons (MLPs), although its roadmap includes natural-language processing based on other model types.

In theory, FPGA tools could enable a developer to compile a model directly to RTL and then synthesize that design. In practice, however, high-level synthesis (HLS) produces suboptimal RTL. Supplying RTL is also problematic, as RTL synthesis takes too long and requires customers to use unfamiliar tools. For inference, Xilinx instead supplies a canned overlay that creates a programmable AI engine compatible with a range of models. Deep AI chose an in-between approach, creating a set of presynthesized overlays with each optimized for a class of models rather than just one. Thus, end customers need only program their card with the image (bit stream) corresponding to their selected model.

The Deep AI software hides the abstraction between the model and the underlying hardware blocks. The company provides sparse pretrained models for ResNet and Yolo networks trained using the ImageNet and Coco data sets. When retraining on a customer data set, its algorithms measure loss and adjust the sparsity level to meet the customer's accuracy requirements.

A customer can simultaneously run inference and retraining on the same card at about half the respective maximum performance of those functions. Xilinx divides its UltraScale+ FPGAs into slices called super logic regions (SLRs). The Alveo U50's FPGA has two SLRs, so one or both are assignable to training or inference. For environments that are always operating, this capability allows a single card to perform retraining in the background while inference handles real-time data. In a factory or retail location with off hours, the entire card can perform retraining during that time. By adjusting its sparsity setting, the customer can tune the retraining to fit the available time, perhaps forfeiting some accuracy but increasing the retraining frequency.

Distance Learning for Machines

The primary alternative for edge training is Nvidia's T4 card, which dissipates 70W (maximum) and provides better performance per watt than the company's new A100 (Ampere). Using the TensorFlow framework and ImageNet data set, Nvidia rates the T4 at 388 images per second (IPS) for ResNet-50 v1.5 training. This example employs mixed precision, which combines FP16 multiplication and FP32 accumulation. The company supplies pretrained models in its Transfer Learning Toolkit, but available performance enhancements are for inference only, so retraining throughput remains the same as training. By comparison, Deep AI rates its software running on the U50 at 2,000 IPS for ResNet-50 v2 using Keras over TensorFlow (tf.keras). Although the

models differ, the U50 delivers 5.2x the T4's throughput; its performance-per-watt advantage is a slightly lower 4.8x.

But Nvidia offers great advantages in flexibility and compatibility. Its Cuda software is the de facto standard for accelerating neural networks, so virtually every model and framework works with Tesla cards, including the T4 (see *MPR 1/27/20,* "Nvidia AI Software Hard to Beat"). By contrast, Deep AI supplies a limited retraining solution targeting specific applications, primarily in image processing and machine vision. For the models it addresses, the company stands out with a combined inference and retraining product well suited to infrastructure-edge systems.

It's still early days for hardware acceleration of such systems, and Deep AI's enterprise customers are likely to favor a hybrid-cloud approach, blending on-premises edge processing and cloud-based services. Microsoft is at the forefront for edge acceleration, offering its Azure Stack Edge Pro appliance with optional T4 cards (up to two) or an Intel Arria 10 FPGA. The latter option handles the same neural-network models as Azure cloud instances, but FPGA acceleration is limited to inference. Presently, Amazon's on-premises AWS Outposts lack hardware-accelerator support.

Such nascent markets are a strong fit for small nimble startups like Deep AI. Because the company needn't fund costly chip development, revenue from only a handful of large customers could allow it to reach profitability and expand its offering into new markets. By focusing on retraining, it delivers unique benefits relative to incumbent Nvidia. If Deep AI's initial product lives up to its claims, silicon partners should line up to enable the startup's roadmap. ♦