November 2020

EDLAB: A BENCHMARK TOOL FOR EDGE DEEP LEARNING ACCELERATORS

HP INC

Title: EDLAB: A Benchmark Tool for Edge Deep Learning Accelerators

Abstract

Plenty of edge deep learning accelerators are proposed to speed up the inference of deep learning algorithms on edge devices. However, Various edge deep learning accelerators feature different characteristics in terms of power and performance, which makes it a very challenging task to compare different accelerators according to their specifications and in turn prohibits a new DL model from being effectively and efficiently deployed on a suitable edge device. We introduce EDLAB, an edge deep learning accelerator benchmark tool, to evaluate the overall performance of edge deep learning accelerators. EDLAB is an end-to-end benchmark tool that provides unified workloads, deployment policy, and fair comparison methodology. Moreover, EDLAB is designed with good scalability, which can support many emerging deep learning applications and hardware.

Detailed Description

Figure 1 presents how EDLAB works as a unified benchmark framework to conduct fair and quantitative evaluation of EDLAs. EDLAB takes as input existing deep neural network (DNN) models. These models are preprocessed by the integrated model conversion tool of EDLAB and then sent to different EDLAs to perform inference. After the inference finishes, we will collect and analyze the data of multiple metrics, such as the inference latency and power consumption, to comprehensively evaluate EDLAs.

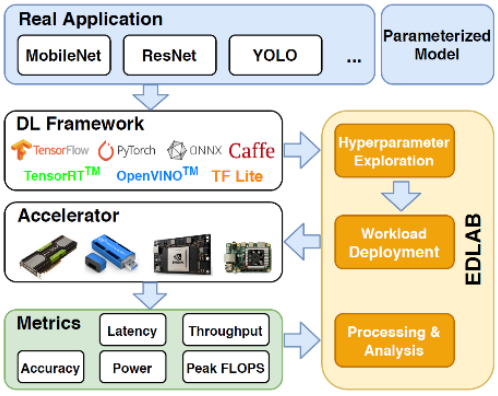

Figure 1. The framework of EDLAB

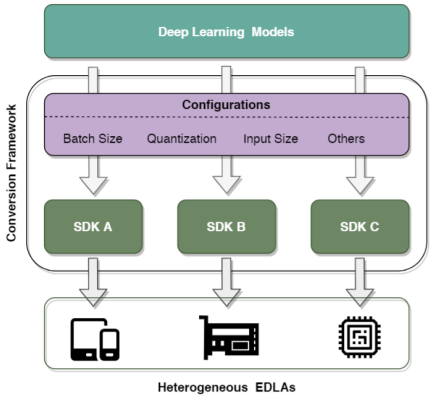

Figure 2. Unified model conversion

A. Benchmark Models

For the current EDLAB, we mainly target EDLAs with resource and power constraints, so we carefully selected several distinct DNN models as shown in TABLE I, which includes models of different magnitudes. In addition, DNN design is gradually shifting from manual design to automatic design, thus we take MnasNet designed by Neural Architecture Search (NAS) to follow the development trend. However, it is worth noting that there is not any limitation in selecting the SOTA models for evaluation in EDLAB. Users are free to select the desired models. We have prepared API to seamlessly add new models into EDLAB.

TABLE I. SOTA DNN models. These models are trained on the training dataset of ILSVRC2012

Model	MACs(G)	Params(M)	Top-1	Description
Inception V3	5.73	27.16	78	Handmade
MobileNet V2	0.3	3.47	71.8	Handmade
MnasNet A1	0.32	3.9	74.5	NAS

B. Workload preprocessing

The successful adoption of EDLAs relies on general DL development frameworks, such as TensorFlow, PyTorch, MXNet, as well as vendor-provided software tool sets. The general frameworks are employed to design and train models while vendor-provided tools optimize the trained models and subsequently convert the optimized models into supported formats. Most vendor-provided tools are incompatible with each other, thereby leading to significant engineering efforts and difficulties for machine learning practitioners who do not have too much knowledge about EDLAs and their tools. Therefore, in EDLAB, we integrate tools provided by different vendors into one unified and integrated framework as illustrated in Figure 2, so that these tools can be easily used to convert different models for diverse EDLAs. The workflow of the conversion framework is that it takes pre-trained models usually obtained from general DL frameworks, and then selects the proper configuration for model conversion and deployment. Finally, it executes the converted workloads on target EDLA for benchmarking.

C. Benchmark Metrics

The design goal of EDLAB is to provide a comprehensive evaluation for EDLAs which can cover various design concerns of edge systems. Therefore, in EDLAB, we evaluate and report several metrics as follows:

Accuracy (Acc): Usually the accuracy can be reproduced once the model structure is determined and weights are well trained and fixed. However, EDLAs employ different tools to convert pre-trained models for efficient execution and such conversion may change the model structures, data representation and bit-width, hence it may cause accuracy drop. It is important to report this accuracy drop for accuracy sensitive applications.

Latency (L): Latency indicates the time elapsed between one input data and its corresponding prediction. For some systems with rigorously temporal requirement, e.g., autonomous driving, latency is essentially critical. In EDLAB, we only measure the latency of the given model processing one image. To avoid the variance caused by different system status, EDLAB runs a model by many times and calculates the average value as the model's latency.

Throughput (T): The throughput of an EDLA is the largest number of samples processed within a time unit, like one second or one minute. The throughput is considered as an important metric for some applications, such as video analytic, which relies on the high throughput performance of hardware to guarantee high Frame-Per-Second (FPS) requirement.

FLOPS (F): We also measure the number of Floating-point Operations executed Per Second (FLOPS) to evaluate the execution efficiency of EDLAs. FLOPS can help to evaluate the execution efficiency of different EDLAs under various DNN models. FLOPS is derived from the throughput and the number of operations which a DNN model has, shown in Eq. 1

$$\boldsymbol{F = T \times FLOPs} \quad \textbf{(1)}$$

where T is the throughput of a DNN model on an EDLA and $FLOPs$ denotes the number of floating-point operations the DNN model has in total. Note that $FLOP(S)$ indicates the hardware performance while $FLOP(s)$ represents the DNN model's complexity

Power ($\boldsymbol{P}$): EDLAB measures the power consumption of EDLAs. Some edge intelligence systems, especially those supplied by batteries, are subject to limited power budget and prefer to achieve an expected performance under limited power budget. However, power measurement relies on the power sensor to accurately obtain it. For those EDLAs which have internal on-board sensors, like Nvidia Jetson series, we can directly obtain power consumption from power sensors. For those without on-board sensors, power must be measured by an external power meter.

Efficiency ($\boldsymbol{E}$): The efficiency of EDLAs is a combined and more comprehensive metric, which is calculated as Equation 2.

$$\boldsymbol{E = \frac{T}{P}} \quad \textbf{(2)}$$

where T is the throughput and P is the power consumption of the accelerator. The unit of efficiency is images per second per watt (imgs/s/W). This metric unifies performance and power into one combined metric, which is instrumental when comparing various EDLAs with different computational capability and power consumption levels.

Disclosed by Liu Weichen, Ravi Subramaniam, Kong Hao, Huai Shuo, Liu Di, Zhang Lei, Chen Hui, Zhu Shien, Li Shiqing, HP Inc.