Proof of concept energy profiler for neural networks

Linus Jungemann
Paderborn University
Dept. Computer Science and *Paderborn Center for Parallel Computing*
Paderborn, Germany
linus.jungemann@uni-paderborn.de

***Abstract*—Neural networks are becoming increasingly important in computer science and industry. At the same time, neural networks and specifically neural network inference are very energy intensive. In this paper, a proof of concept neural network energy profiler for neural networks in datacenters is presented. Additionally, the implemented neural network energy profiler is used to profile the energy efficiency of an example neural network on CPUs, GPUs and FPGAs.**

***Index Terms*—Neural Networks, AI, FPGA, FINN, Energy Efficiency, Inference**

I. Introduction

(Deep) Neural networks have become one of the most prevailing topics in computer science and industry alike. It is likely that the amount of users of applications built using neural networks will increase significantly in the next couple of years as more and more companies invest into artificial intelligence which is powered by neural networks [1].

Due to the working principles of deep neural networks, it is very resource intensive in both time and computational resources to train a deep neural network. It is estimated that training of one energy intensive neural network can cost as much as $656\,347\,\mathrm{kW\,h}$ of energy and therefore result in $284\,\mathrm{t}$ of CO_2 emitted to the atmosphere [2]. At the same time, the size of neural networks is increasing quickly.

Although no peer reviewed studies are currently available, it is estimated that the inference phase surpasses training of neural networks in terms of computational costs and therefore energy intensity by $8 - 9$ times [3]. The main reason for the large impact of the inference phase on the energy intensity of neural networks, is that the inference operation is executed significantly more often than the training phase during the lifetime of a neural network.

Currently, mostly Central Processing Units (CPUs) and Graphics Processing Units (GPUs) are used for inference computation. Both, CPUs and GPUs face the problem of a relatively low energy-efficiency due to their general purpose architecture. One solution for the issue of energy-efficiency in the task of inference for neural networks could be Field Programmable Gate Arrays (FPGAs). Due to their ability to be programmed to provide a hardware architecture that is specialized for the inference of neural networks by implementing them in hardware, it could be possible to reduce the power consumption and therefore CO_2 emissions of inference significantly.

Several solutions to map (deep) neural networks to FPGAs exist, such as FINN [4] and fpgaConvNet [5]. At the same time, only a few publications research the energy efficiency of deep neural networks inference on FPGAs and to my knowledge, there are no general tools to measure the power consumption/efficiency of such mapped deep neural networks.

Therefore, the proof of concept for a new extensible tool will be presented in this paper to profile the energy efficiency of neural networks. The implemented tool is independent of the type of neural network to be profiled, and independent of the hardware used for execution.

The proof of concept profiler will be used to evaluate the energy efficiency of neural network inference on CPU, GPU and FPGA hardware for an example network.

II. Approach

A modern datacenter server consists of one or more Central Processing Units (CPUs) and optionally one or more Graphics Processing Units (GPUs) or Field Programmable Gate Arrays (FPGAs). Nowadays, these components provide interfaces to be able to measure the momentary power consumption of these components.

Modern datacenter CPUs implement the Intel's Running Average Power Limit (RAPL) interface, which allows a user to query the power consumption of a single core and package power for each processor in the system. The measurement values are based on hardware performance counters included in the CPUs.

Accelerators such as GPUs and FPGAs usually include capabilities to read the momentary power consumption reported by sensors based on the accelerator cards itself. The data is either provided by a vendor specific library such as `nvml` by NVIDIA or by the `sysfs` file system of the operating system.

Additionally, server systems nowadays often implement the Intelligent Platform Management Interface (IPMI), which can be used to access the power consumption of the complete server.

The implemented proof of concept energy profiler uses these system interfaces and records each data source individually for later processing. The neural network under test is controlled by the energy profiler during the profiling process using a user provided interface to the network under test.

III. Results

Evaluation of the proof of concept energy profiler was conducted using the Noctua 2 cluster at Paderborn Center for

Node Type	CPU Node	GPU Node	FPGA Node
CPU	2x AMD Milan 7763		2x AMD Milan 7713
RAM	256 GB	512 GB	
Accelerator	None	Nvidia A100	AMD Xilinx Alveo U280

TABLE I

Overview of the three node types (CPU Node, GPU Node and FPGA Node) and corresponding hardware used for evaluating the implemented neural network profiler.

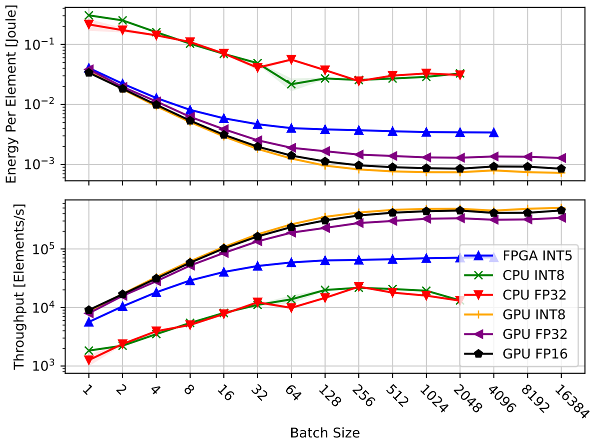

Fig. 1. Top graph: Energy consumed per inference operation for a single input element over different batch sizes for the RadioML VGG10 model. Bottom graph: Throughput (elements per second) for the same model for different batch sizes. Both graphs include results for the CPUs, the GPU and the FPGA and the respective bit widths tested on them. Energy consumed by miscellaneous components (mainboard, memory, network card) is excluded.

Parallel Computing. An overview of the hardware used for evaluating the energy profiler can be found in Table I. As an example for profiling the energy consumption, the RadioML VGG10 neural network presented by Jentzsch et al. [6] was selected. The network was selected, as it was already successfully deployed to FPGAs using the FINN framework, which was also utilised in this work. The deployment to CPU was handled using Apache TVM, while TensorRT was employed for deployment to GPU. On CPU and GPU, both inference using the FP32 and INT8 datatype were tested. Additionally, FP16 was tested on GPU only, as it is not supported on CPUs at the moment. For the FPGA, the inference was conducted using 5-bit integers, the largest datatype supported by the FINN framework at the moment.

The results (as displayed in Fig. 1) show that in terms of throughput (bottom graph), the FPGA implementation is located between the GPU and CPU implementations. The FPGA implementation has a significantly higher throughput than the CPU implementation (FP32, as well as INT8). Compared to the GPU implementation, the inference using FPGAs results in a lower throughput for all input batch sizes and data types used for GPU inference. This difference is less pronounced for smaller input batch sizes.

A similar effect can be seen when looking at the energy consumed by each element during inference (Fig. 1, top graph). Again, CPU inference (INT8, FP32) is less efficient compared to FPGA inference, which in turn is less efficient compared to the inference using a GPU (INT8, FP16 and FP32). The difference in energy efficiency between GPU and FPGA inference is less pronounced for smaller batch sizes.

These results do not seem to match with the idea of implementing neural network inference using FPGAs to enable higher efficiency inference. When analysing the results of the inference using FPGAs in more detail, it is possible to see that 40-99% of the FPGA execution (depending on the input batch size) is spent in the FPGA driver used by the FINN framework. This inefficient implementation results in a big overhead in inference latency, throughput and energy efficiency and likely results in the significantly worse energy efficiency when implementing inference using FPGAs compared to GPUs.

IV. Outlook

Based on these preliminary results, there exist some clear areas of future work. First, the proof of concept implementation of the neural network energy profiler will be extended into a full, easy to use open-source tool. Furthermore, a new driver for the FINN framework is currently in development, which has a considerable lower overhead compared to the current implementation to increase the energy efficiency of neural network inference using FPGAs.

An additional area of planned research is the prediction of the energy consumption and efficiency for neural networks on different sets of hardware. Synthesis of neural networks for FPGAs takes a long time and a large amount of energy. Therefore, it is vital to have a tool that allows to predict the energy consumption of a neural network on different hardware architectures and can give out estimations about the best hardware to be used for inference deployment.

References

[1] L. Columbus. 76% of enterprises prioritize AI & machine learning in 2021 IT budgets. Section: Enterprise Tech. [Online]. Available: https://www.forbes.com/sites/louiscolumbus/2021/01/17/76-of-enterprises-prioritize-ai–machine-learning-in-2021-it-budgets/

[2] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in NLP," *Computing Research Repository*, no. arXiv:1906.02243, pp. 1–6, Nov 2019. [Online]. Available: https://doi.org/10.48550/arXiv.1906.02243

[3] Moor Insights and Strategy. Google cloud doubles down on NVIDIA GPUs for inference. Section: Enterprise & Cloud. [Online]. Available: https://www.forbes.com/sites/moorinsights/2019/05/09/google-cloud-doubles-down-on-nvidia-gpus-for-inference/

[4] Y. Umuroglu, N. J. Fraser, G. Gambardella, M. Blott, P. Leong, M. Jahre, and K. Vissers, "FINN: A framework for fast, scalable binarized neural network inference," in *Proc. Int. Symp. on Field-Programmable Gate Arrays (FPGA)*. Monterey, California, USA: ACM, Feb 2017, pp. 65–74. [Online]. Available: https://doi.org/10.1145/3020078.3021744

[5] S. I. Venieris and C.-S. Bouganis, "fpgaConvNet: A framework for mapping convolutional neural networks on FPGAs," in *Proc. Int. Symp. on Field-Programmable Custom Computing Machines (FCCM)*. Washington, DC, USA: IEEE, May 2016, pp. 40–47. [Online]. Available: https://doi.org/10.1109/FCCM.2016.22

[6] F. Jentzsch, Y. Umuroglu, A. Pappalardo, M. Blott, and M. Platzner, "RadioML meets FINN: Enabling future RF applications with FPGA streaming architectures," *IEEE Micro*, vol. 42, no. 6, pp. 125–133, Nov 2022. [Online]. Available: https://doi.org/10.1109/MM.2022.3202091