Are CPUs good for energy-efficient deep learning ?

Master internship

Expected starting: February/March 2023 (6 months)

Keywords

Deep learning, CPU, Optimized Neural Networks

Context and objectives

It is recently that Deep Learning (DL) has become inseparable from GPUs/TPUs, allowing parallel computing to be carried out more efficiently. However it can be difficult for many users to adapt their hardware to use DL-based solutions, reducing the adoption of such high-performance solutions.

Hence we saw the advancement of embedded Deep Learning, where smaller GPUs/TPUs with lower power consumption are used. The most striking example is surely the use of GPUs integrated into smartphone processors.

Therefore, to operate on such platforms, we started to see the use of techniques such as pruning (reducing the number of weights of a neural network) or quantization (reducing the number of bits) and we can then wonder if it is still necessary to have a GPU in the case of inference on pruned and quantized networks.

The objective of this internship is to assess performance of DL inference for a given task according to different configurations, GPU/CPU, with/without pruning [4] and with/without quantization [3]. The focus will not be limited to execution times and latency but also include power consumption as well. All methods considered in the study will be off-the-shelf methods, allowing effort not to be allocated solely to methodological development.

More specifically, this work will focus on bringing potential optimization [2] or use of libraries [1] that could allow DL to help other research communities without having the financial and energetical burden that is often associated.

Potential outcomes of the internship will lead to publications in remote sensing, computer vision or machine learning fields, depending on the nature of the contributions.

Work program

In order to address the aforementioned objectives, a tentative work program is given below.

- Bibliographical study of techniques involving Deep Learning and CPUs and network optimizations.
- Benchmark of a network for a given CPU/GPU task.
- Pruning/Quantization of this network using methods available on public libraries.
- Benchmark of the different configurations on CPU/GPU.

Required background and skills

- Master 2 Student or equivalent with an excellent academic track;
- Background in computer science and/or machine/statistical learning and/or applied mathematics for signal and image processing;
- Excellent programming skills in Python (familiar with at least one deep learning package, such as TensorFlow, PyTorch or JAX is a must.)

Supervision

The expected intern will join the OBELIX research group (`www.irisa.fr/obelix`) from IRISA (UMR 6074) that is located in the UBS (Université Bretagne Sud) campus in Vannes 56000, France.

The internship will be jointly supervised by **Jean-Christophe Burnel** (PhD student at UBS) and **Prof. Sebastien Lefevre** (Professor at UBS).

Application

Send your CV + Motivation letter + Master transcripts to `jean-christophe.burnel@irisa.fr` and `sebastien.lefevre@irisa.fr` **(as soon as possible and before January 30, 2023)**. Potential candidates will be contacted for interview.

References

[1] Riel D Castro-Zunti, Juan Yépez, and Seok-Bum Ko. License plate segmentation and recognition system using deep learning and OpenVINO. *IET Intelligent Transport Systems*, 14(2):119–126, 2020.

[2] Sparsh Mittal, Poonam Rajput, and Sreenivas Subramoney. A survey of deep learning on CPUs: opportunities and co-optimizations. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[3] Eriko Nurvitadhi, David Sheffield, Jaewoong Sim, Asit Mishra, Ganesh Venkatesh, and Debbie Marr. Accelerating binarized neural networks: Comparison of FPGA, CPU, GPU, and ASIC. In *2016 International Conference on Field-Programmable Technology (FPT)*, pages 77–84. IEEE, 2016.

[4] Michael Zhu and Suyog Gupta. To prune, or not to prune: exploring the efficacy of pruning for model compression. *arXiv preprint arXiv:1710.01878*, 2017.