Ph.D. Thesis: Embedded Deep Learning for Real-Time Vision Applications

Keywords: convolution, deep learning, convolutional neural network, embedded system, real-time stream processing, FPGA, HLS, VHDL

Objectives

Deep learning has become of significant interest from the researchers and manufacturer, mainly because of its higher performances than classical methods. However, for applications in Edge Computing, IoT, and Embedded Systems (autonomous driving, drones, etc.), their high computational demands and the low energy-efficiency of GPUs for such applications have led to consider custom hardware accelerators. As a consequence, the use of FPGA has raised interests by the community thanks to their modularity and energy-efficiency [1]. In this direction, lots of recent works have shown that accelerating the inference of convolutional neural networks (CNNs) using FPGAs could be made efficient for embedded systems, sometimes even very close to their original GPU implementation [2], while having the energy-efficiency from FPGAs.

For all of these studies, these performance evaluations are biased towards the standard evaluation pipeline of a model. Indeed, all data are known before starting the processing, they are all stored in memory, many of them could be computed together by using a batch of images as input (the same operation are performed on each image), etc. As such, too many credits are given to methods with high throughput and too few to methods with small latency, which, among others, is a critical criterion in all real-time applications [3]. As a consequence, this evaluation process does not cover most of the manufacturer needs, such as data streaming (e.g., real-time video processing in autonomous driving), single image processing (e.g., a call to an API to search a given object in a database of product), etc.

The current experimental setup is also very limited: experiments are limited to small networks (typically AlexNet [4] with 11 layers) which are not representative to the current state-of-the-art architectures (e.g., ResNet [5] with 50 to 152 layers or DenseNet [6] with up to 269 layers). Also, they are limited to the image classification task, letting unexplored tasks that could benefit from a real-time implementation (e.g., for autonomous vehicles) such as object detection with R-CNN architectures [7], or semantic segmentation with U-Net architectures [8].

In this thesis, we target the category of applications where data are streams, real-time is required, and networks are deeper. These requirements mean that our FPGA implementations of CNNs have to be efficiently parallelized to address the complex computing pipeline of data streams, and the temporal constraints of real-time applications (latency, flows of data, etc.) while keeping the following criteria:

- **Flexibility**: to allow their use on multiple applications (different CNN architectures or tasks, other FPGA devices, etc.)

- **Deterministic**: for the control of computation, communication times to respect the real-time constraint
- **High performance**: to keep performances close to their GPU implementation counterparts

Main Research Axes

The primary objectives from this thesis will be divided into the following axes:

- **Performance Evaluation of state-of-the-art methods for real-time streaming**
  - Building a benchmark for real-time data streaming computer vision applications (which applications and which metrics)
  - Benchmarking state-of-the-art methods on this task
- **New Parallelized Implementations**
  - Improving current strategies to deal with real-time data streaming CNN (parallelization, dataflow, heavily pipelined architectures, etc.)
  - Study the impact of data type (floating point, fixed point, integer, etc.)
  - Exploit sparsity in CNNs

Desired Skills

- Embedded systems and real-time applications
- Experiences with device design software: Quartus or Vivado
- Experiences in HLS
- Languages: VHDL/Verilog, C/C++
- Knowledge in deep learning and CNN is not mandatory
- Spoken and written English are mandatory

Degrees

The candidate is expected to have (or in the process of finishing) a Master degree in Computer Science or in Electronic Engineering. A Research Master in Embedded Systems or a first experience in a research environment would be a plus.

Funds and Collaborations

This thesis is expected to be fund by a Franco-Singaporean collaboration for 4 years. The candidate will work in ETIS lab (Cergy-Pontoise, France) during 2 years, and the two others will be spent at Singapore. A tight collaboration will be held during these 4 years.

Supervisors

- Mohamed Amine KHELIF, ETIS lab, mohamed-amine.khelif@ensea.fr
- Pierre JACOB, CTU Prague (Czech Republic), jacobpie@fel.cvut.cz
- Aymeric HISTACE, ETIS lab, aymeric.histace@ensea.fr

How to Apply

Applicants should submit a curriculum vitae and a cover letter in English to all supervisors listed above.

References

[1] Y. Chen, T. Krishna, J. S. Emer, and V. Sze. Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks. *IEEE Journal of Solid-State Circuits*, Jan 2017.

[2] Kamel Abdelouahab, Maxime Pelcat, Jocelyn Serot, and François Berry. Accelerating cnn inference on fpgas: A survey. *arXiv preprint arXiv:1806.01683*, 2018.

[3] A. A. Gilan, M. Emad, and B. Alizadeh. Fpga-based implementation of a real-time object recognition system using convolutional neural network. *IEEE Transactions on Circuits and Systems II: Express Briefs*, 2019.

[4] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[6] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[7] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[8] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.