Special Issue on Embedded Vision Architectures for Machine Learning

Francois Berry¹ · Luca Maggiani² · Ricardo Carmona-Galán³

Accepted: 22 December 2021 / Published online: 10 January 2022

© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2022

Machine learning techniques, particularly those based on deep neural networks, are widely used for many artificial intelligence applications—including computer vision, speech recognition, robotics, etc.—as they offer cutting-edge accuracy. This performance is achieved, however, at the cost of a high computational complexity, what becomes a severe limitation for their implementation in resource-constrained autonomous embedded vision systems. The challenge is to reduce the amount of computation initially required to implement complex deep neural networks, to fit them into heavily constrained low-resource platforms. This problem can be approached at different levels, from the simplification and reduction of the network elements, in order to implement an approximated solution, to the design of highly adapted signal processing architectures.

This special issue presents different architectural compromises involved in the execution of heavy computational loads related to machine learning on CPUs, GPUs, FPGAs and ASICs, and even their impact on memory hierarchies and storage. This subject has always been a central topic at the Workshop on Architecture of Smart Cameras (WASC), a workshop especially dedicated to bring together researchers and engineers covering the all aspects of the implementation of smart cameras. It has been held in different cities, from Clermont-Ferrand in 2012 to Ghent in 2020. Smart cameras are embedded vision systems that are required to

produce a semantic understanding of the scene and to generate a response. The incorporation of deep neural networks represents a significant leap in performance, but an efficient implementation is needed not to compromise the scarce resources found in embedded platforms. This special issue focuses on addressing this issue.

The first paper is dedicated to the joint adaptation of hardware and software to facilitate the integration of convolutional neural networks into embedded systems. Lucas Fernandez-Brillet and his colleagues at the Laboratoire Techniques de l'Informatique et de la Microelectronique pour l'Architecture de Systemes Integres (Grenoble, France), propose a new general CNN compression method to reduce the number of parameters and operations, based on principal components analysis. They show overall compression rates of $2\times$ can be achieved on a compact ResNet-32 model on the CIFAR-10 dataset, with only a negligible loss of 2% of the network accuracy, while up to $11\times$ compression rates can be achieved on specific layers with negligible accuracy loss.

A second paper, from Luis Rodriguez-Benitez and his colleagues from the Universidad de Castilla-La Mancha (Spain), deals with a new paradigm called the Internet of Agents (IoA), which allows the inclusion of intelligence and autonomy on IoT devices and networks. In order to achieve that, they have designed specific protocols for the transmission of models and the training set. The convolutional neural network is trained in the cloud, then simplified and passed to agents located at the embedded systems to implement inferences.

A third paper implements a reasoning model based on a Bayesian network into a low cost computer vision system, which is able to generate traffic metrics at urban intersections. This paper is authored by Gines Domenech-Asensi and his team at the Universidad Politecnica de Cartagena (Spain). This system works on data obtained from background subtraction and contrast analysis techniques. It has been prototyped on a low-cost platform and also deployed to surveillance video cameras managed by the local traffic and transport administration.

✉ Francois Berry
francois.berry@uca.fr

Luca Maggiani
luca@sma-rty.com

Ricardo Carmona-Galán
rcarmona@imse-cnm.csic.es

¹ UMR 6602, Institut Pascal - CNRS, Avenue Blaise Pascal, 63178 Aubière, France

² Sma-RTy Italia SRL, Via Felice Cavallotti 133, 20093 Milano, Cologno Monzese, Italy

³ Instituto de Microelectrónica de Sevilla (CSIC-Universidad de Sevilla), C/Americo Vespucio 28, 41092 Sevilla, Spain

Finally, a paper from Zhang Xianhong and colleagues from the Northeast Forestry University, (Harbin, China) is dedicated to the application of generative adversarial networks to the enhancement of remotely sensed images. These images are usually affected by colour shift, degraded textures information and, in general, poor quality. The authors have approached the low efficiency and the distortion of image details with the implementation of a GAN based on texture attention.

In summary, we are presenting here several approaches to the efficient implementation of complex networks in

low-resource embedded platforms. Examples with commonly found image recognition and object detection problems are also presented. It is our hope that these proposals help the community to continue progressing towards flexible enough but still efficient solutions for the architecture of smart cameras.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.