Neural Networks and Deep Learning: Part 3: Implementation issues

The first lecture is scheduled on March 15, 2022 - 9:00
Grey Room, Retis Lab, TeCIP Institute, Via Moruzzi, 1 – Pisa

Giorgio Buttazzo and Alessandro Biondi

Objectives: The aim of the part is to discuss practical and implementation issues useful to deploy neural networks on a variety of embedded platforms using different languages and development environments.

Course program

1. Implementing Neural Networks from scratch in C. General implementation principles. Main and auxiliary functions.
2. Sample implementations of common neural network models in C language.
4. Functional components on autonomous driving. Basic blocks for perception, prediction, planning, control, and actuation.
7. DNN optimization for embedded platforms. Weight quantization, pruning, distillation, to reduce execution times and contain memory footprints in resource constrained platforms.
8. Accelerating deep networks on GPGPUs. Overview of the Nvidia TensorRT framework. Executing a DNN modelled in Caffe in TensorRT.
9. GPU-based real-time neural vision. How to accelerate a neural network on TensorRT to detect objects from a video camera.
10. Accelerating deep networks on FPGA. Technologies and approaches.
11. Design and optimization of DNNs accelerators on FPGA.