Accelerating Deep Neural Networks on Mobile Processor with Embedded Programmable Logic

e-Lab

Aysegul Dundar*, Jonghoon Jin*, Vinayak Gokhale*, Bharadwaj Krishnamurthy, Alfredo Canziani, Berin Martini and Eugenio Culurciello

(*contributed equally)

PURDUE UNIVERSITY

Abstract

We present a mobile platform aimed at accelerating deep convolutional neural networks (DCNN). DCNN is a powerful way to categorize images. They have achieved state of the art performance in many visual classification benchmarks. Their computational costs prevent them from being deployed for real-time applications.

We implemented a hardware accelerator on a logic device (Xilinx Zynq) in order to run DCNNs in real-time. The platform consists of programmable logic (FPGA) and a mobile CPU (ARM Cortex-A9 2x cores), sharing the same memory (DDR3).

The accelerator allows ~25x faster execution than on CPU alone.

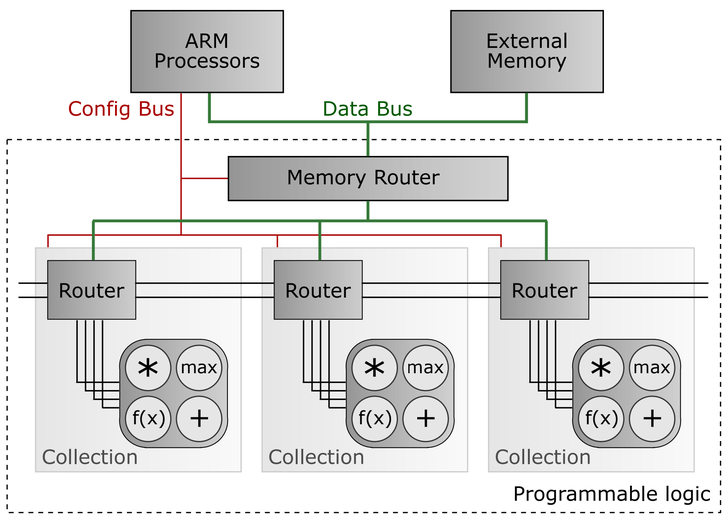

Fig.1. Diagram of the system and streaming paths. Neighboring connections between collections construct a torus-like data streaming network

Performance & Demonstration

Performance Comparison

	Intel i7 4-core	nVidia GTX780	nVidia GT650m	CPU ARM 2-core	nn-X Zynq	nn-X 4x (ZC706)
Peak (GOP/sec)	200	3800	182	10	40	120
Actual (GOP/sec)	90	620	54	0.4	36	100
Power (W)	45	500	30	2	3.9	5
Embeddable factor (GOP/s/W)	2	1.24	1.8	0.2	9.231	20

Table.1. Performance and power consumption computed on a 16x10x10 filter-bank over a 500x500 input image

Filter Bank

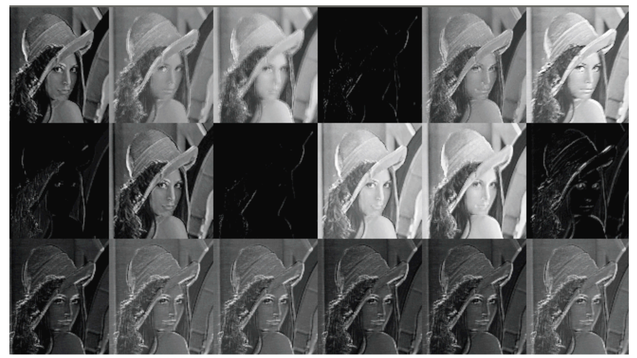

Fig.2. Full-convolution-loop with 18x10x10 filters over a 500x500 image, ~50x faster than CPU

Face Detector

Fig.3. Face detector on 400x400 multi-scale image ~10x faster than CPU

Street-scene Parsing

Fig.4. Street-scene object categorization and labeling. Performs full-scene understanding with 10 categories on a 320x200 video ~30x faster than CPU