Direct Convolution: Performance Effects of Loops Ordering and Parallelization

Mirco Manninona, Andrea Mondellib, Sandro Bartolinić

*a Dipartimento di Ingegneria dell’Informazione e Scienze Matematiche, Via Roma 56, 53100 Siena, Italy

b Huawei Technologies Research Development (UK) Ltd, Cambridge, United Kingdom

e-mail: mannino@dism.unisi.it

BACKGROUND

• Direct convolution is a technique used to address convolution that can be implemented through (at least) six perfect nested loops surrounding an accumulation operation over output elements.

• Advantages over indirect convolution methods (i.e., methods that apply transformations to tensors) [1]:
  • No memory overhead.
  • Better scaling as the number of threads increases.
  • HW/SW codesign is needed to design convolutional accelerators that guarantee high performance and low power consumption.

LOOP ORDERINGS

Order N.1

Order N.2

Order N.3

The dimension used to parallelize is the output height (Ho). Each thread has been assigned a portion of the output rows to be processed. It is possible to assign a portion of the output to each thread because every output element is completely independent from others.

Scalability

Order N.1

Order N.2

Order N.3

Cache usage

In AlexNet, kernel tensor has the highest memory footprint.

Using up to 4 threads, input and output memory footprint decreases to L2 threshold. This is reflected in the doubling of performance.

Using 8 threads, memory footprint of output and input further decreases, but performance are limited to kernel tensor (stored in L3 shared cache).

FUTURE WORK

• Use multiple degrees of parallelization.

• Investigation of other memory layouts for scalability improvements.

• Definition of ad-hoc computational and memorization units in RISC-V accelerator, based on results of this work.

• We are investigating general instructions for handling RISC-V accelerators. This work will be of crucial importance to test them.

PLATFORM

• Results obtained with gem5 simulator [3] using RISC-V ISA in Syscall Emulation (SE) mode.

• Cache configuration:
  • L1: 32 KB
  • L2: 256 KB
  • L3 (shared): 16 MB

