

CONVOLUTION ACCELERATION AND COMPUTING PROCESSING METHODS AND APPARATUSES, ELECTRONIC DEVICE, AND STORAGE MEDIUM

5

ABSTRACT

Convolution acceleration and computing processing methods and apparatuses,
electronic device, and a storage medium. The acceleration method comprises: reading
10 data to be processed of a preset size in a task to be processed from an off-chip
memory of an accelerator in sequence by means of a FIFO port, and storing same in at
least one input cache region of a first cache region in an on-chip memory of the
accelerator; in response to storage of the data to be processed in the input cache
region, reading input window data to be processed from at least one input cache
15 region in sequence, and performing convolution calculation on the input window data
to be processed by means of convolution kernel data to obtain output window data,
and storing the output window data in a third cache region; and storing the output
window data in the third cache region in the off-chip memory in sequence by means
of the FIFO port. The convolution acceleration and computing processing methods
20 and apparatuses, the electronic device, and the storage medium can avoid the cases
that the on-chip memory of the accelerator and bandwidth are insufficient and the
processor resources are limited, improve the efficiency of convolution processing, and
are appropriate for hardware platforms such as FPGA and ASIC.

25

FIG. 1