

PERFORMING AVERAGE POOLING IN HARDWARE

ABSTRACT

Methods and systems for receiving a request to implement a neural network comprising an average pooling layer on a hardware circuit, and in response, generating instructions that when executed by the hardware circuit, cause the hardware circuit to, during processing of a network input by the neural network, generate a layer output tensor that is equivalent to an output of the average pooling neural network layer by performing a convolution of an input tensor to the average pooling neural network layer and a kernel with a size equal to a window of the average pooling neural network layer and composed of elements that are each an identity matrix to generate a first tensor, and performing operations to cause each element of the first tensor to be divided by a number of elements in the window of the average pooling neural network layer to generate an initial output tensor.

FIG. 7