PERFORMING KERNEL STRIDING IN HARDWARE

ABSTRACT

Methods for receiving a request to process, on a hardware circuit, a neural network comprising a first convolutional neural network layer having a stride greater than one, and in response, generating instructions that cause the hardware circuit to, during processing of an input tensor, generate a layer output tensor equivalent to an output of the first convolutional neural network layer by processing the input tensor using a second convolutional neural network layer having a stride equal to one but that is otherwise equivalent to the first convolutional neural network layer to generate a first tensor, zeroing out elements of the first tensor that would not have been generated if the second convolutional neural network layer had the stride of the first convolutional neural network layer to generate a second tensor, and performing max pooling on the second tensor to generate the layer output tensor.

[Fig. 9]