

## CLAIMS

1. A hardware circuit configured to process a convolutional neural network layer having a particular stride that is greater than one, the hardware circuit comprising:

matrix computation circuitry configured to:

generate a first tensor by convolving an input tensor received by the hardware circuit and a kernel, wherein the convolution is performed with a stride equal to one, and wherein the first tensor includes (i) each element of a plurality of elements that would have been generated if the convolution had been performed with the particular stride and (ii) additional elements that would not have been generated if the convolution had been performed with the particular stride; and

generate a second tensor by performing element-wise multiplication of the first tensor and a masking tensor to zero out the additional elements that would not have been generated if the convolution had been performed with the particular stride; and

pooling circuitry configured to generate a third tensor corresponding to an output of the convolutional neural network layer having the particular stride by performing max pooling on the second tensor to remove the zeroed out additional elements from the second tensor.

2. The hardware circuit of claim 1, wherein the masking tensor comprises zeros at each position that corresponds to an element of the first tensor that would not have been generated if performing a convolution with the stride greater than one, and ones at each other position of the masking tensor.

3. The hardware circuit of claim 1, wherein the input tensor, the kernel, and the masking tensor are each stored at a memory accessible by the hardware circuit.

4. The hardware circuit of claim 1, wherein performing max pooling on the second tensor comprises: obtaining, for each of one or more windows of the second tensor that are defined by the stride greater than one, a maximum value element from among elements in the window.

5. The hardware circuit of claim 4, wherein the one or more windows of the second tensor that are defined by the stride greater than one have rectangular dimensions, each dimension being equal to the stride greater than one, and each window including different elements of the second tensor.

6. The hardware circuit of claim 1, wherein the convolutional neural network layer having the stride greater than one is a first layer in a neural network, and the input tensor is a representation of a digital image comprising elements that correspond to pixels of the digital image.

7. The hardware circuit of claim 1, wherein the pooling circuitry comprises:

registers configured to receive elements of the second tensor; and

aggregation circuitry configured to access the elements of the second tensor stored at the registers and apply a maximum determining function to the elements of the second tensor stored at the registers.

8. The hardware circuit of claim 7, wherein the pooling circuitry comprises one or more memory units configured to: receive elements of the second tensor; and provide elements of the second tensor to the registers.

9. The hardware circuit of claim 1, wherein the hardware circuit comprises a host interface configured to receive instructions for controlling the hardware circuit to process the convolutional neural network layer having the stride greater than one.

10. The hardware circuit of claim 9, wherein the hardware circuit comprises a sequencer configured to receive the instructions and to generate, based at least on the instructions, control signals for controlling the matrix computation circuitry and the pooling circuitry to process the convolutional neural network layer having the stride greater than one.

11. The hardware circuit of claim 9, wherein the instructions are generated in response to a request to process an input using a neural network that includes the convolutional neural network layer having the stride greater than one.

12. The hardware circuit of claim 1, wherein the input tensor is stored at a unified buffer of the hardware circuit and the kernel is stored at a dynamic memory of the hardware circuit, and wherein the hardware circuit is configured to enable the matrix computation circuitry to generate the first tensor by convolving the input tensor and the kernel by: sending the input tensor from the unified buffer to the matrix computation circuitry; and sending the kernel from the dynamic memory to the matrix computation circuitry.

13. The hardware circuit of claim 1, wherein, to generate the first tensor by convolving the input tensor and the kernel, the matrix computation circuitry is configured to:

- receive elements of the input tensor at one or more cells of the matrix computation circuitry;

- receive weights of the kernel at the one or more cells of the matrix computation circuitry;

- process the received elements of the input tensor and the received weights of the kernel at the one or more cells of the matrix computation circuitry; and output results of the processing to one or more accumulators of the matrix computation circuitry.

14. The hardware circuit of claim 13, wherein each cell of the matrix computation circuitry comprises:

- an activation register configured to receive an element of the input tensor;

- a weight register configured to receive a weight of the kernel; and

- multiplication circuitry configured to multiply the element of the input tensor and the weight of the kernel.

15. The hardware circuit of claim 1, wherein, to generate the second tensor by performing element-wise multiplication of the first tensor and the masking tensor, the matrix computation circuitry is configured to:

- generate a first output tensor by performing matrix multiplication of the first tensor and an identity matrix;

- generate a second output tensor by performing matrix multiplication of the masking tensor and an identity matrix; and

- obtain the second tensor by multiplying the first output tensor and the second output tensor.

16. The hardware circuit of claim 15, wherein the matrix computation circuitry comprises multiplication circuitry configured to multiply the first output tensor and the second output tensor.

17. The hardware circuit of claim 1, wherein the pooling circuitry is separate from the matrix computation circuitry.

18. The hardware circuit of claim 1, wherein:

the convolutional neural network layer having the particular stride is configured to generate an output tensor that has a first number of elements;

the first tensor has a second number of elements that is greater than the first number of elements; the second tensor has the second number of elements; and

the pooling reduces the second number of elements to the first number of elements by removing the zeroed out additional elements.