

NEURAL NETWORK INSTRUCTION SET ARCHITECTURE

BACKGROUND

[0001] This specification relates to an instruction set for computation of Deep Neural Network (“DNN”).

[0002] Neural networks are machine learning models that employ one or more layers of models to generate an output, e.g., a classification, for a received input. Some neural networks include one or more hidden layers in addition to an output layer. The output of each hidden layer is used as input to the next layer in the network, i.e., the next hidden layer or the output layer of the network. Each layer of the network generates an output from a received input in accordance with current values of a respective set of parameters.

[0003] Some neural networks include one or more convolutional neural network layers. Each convolutional neural network layer has an associated set of kernels. Each kernel includes values established by a neural network model created by a user. In some implementations, kernels identify particular image contours, shapes, or colors. Kernels can be represented as a matrix structure of weight inputs. Each convolutional layer can also process a set of activation inputs. The set of activation inputs can also be represented as a matrix structure.

SUMMARY

One innovative aspect of the subject matter described in this specification can be embodied in a computer-implemented method for accelerating inference computations for a neural network having multiple neural network layers and using a system comprising one or more hardware compute units. The method includes receiving, by a processing unit of a hardware compute unit, a single instruction that specifies parameters that: encode a neural network layer type of a neural network layer of the plurality of neural network layers; and are used by the hardware compute unit to perform a tensor computation for the neural network layer; determining, by the processing unit and based on the single instruction, dimensions of at least one multi-dimensional tensor, where the at least one multi-dimensional tensor includes elements that correspond to an operand for performing the tensor computation; generating, by the hardware compute unit and based on the single instruction, a loop nest comprising a plurality of loops, wherein a structure of the loop nest is defined using the neural network layer type indicated by the data values of the single instruction; and

performing, by the hardware compute unit, the tensor computation by executing the loop nest to access operands corresponding to elements of the at least one multi-dimensional tensor. These and other implementations can each optionally include one or more of the following features. For example, the tensor computation can be at least a portion of a computation of a neural network layer. The parameters specified by the instruction may comprise a parameter that specifies a type of the neural network layer, and the structure of the loop nest can be defined at least in part by the type of the neural network layer. Thereby, executing a loop nest comprising a plurality of loops may refer to performing a traversal over elements of a tensor in an order specified by a structure of nested loops, wherein said structure is specified e.g. by a depth of the loop nesting, and start and end indices, the stride and the direction of each loop respectively.

CLAIMS

1. A computer-implemented method for accelerating inference computations for a neural network having multiple neural network layers and using a system comprising one or more hardware compute units, the method comprising:

receiving, by a processing unit of a hardware compute unit, a single instruction that specifies data values that:

encode a neural network layer type of a neural network layer of the plurality of neural network layers; and are used by the hardware compute unit to perform a tensor computation for the neural network layer;

determining, by the processing unit and based on the single instruction, dimensions of at least one multi-dimensional tensor, where the at least one multi-dimensional tensor includes elements that correspond to an operand for performing the tensor computation;

generating, by the hardware compute unit and based on the single instruction, a loop nest comprising a plurality of loops, wherein a structure of the loop nest is defined using the neural network layer type indicated by the data values of the single instruction; and

performing, by the hardware compute unit, the tensor computation by executing the loop nest to access operands corresponding to elements of the at least one multi-dimensional tensor.

2. The method of claim 1, wherein the tensor computation is at least a portion of a computation of a neural network layer.

3. The method of claim 1, wherein the single instruction causes the processing unit to access at least one memory address of an array in a storage medium, the memory address of the array comprising a variable that is read by the processing unit during performance of the tensor computation.

4. The method of claim 3, wherein performing the tensor computation comprises, providing, by the processing unit, at least one control signal to a tensor traversal unit (TTU) to cause the TTU to emit loop indices used in executing the loop nest during performance of the tensor computation.

5. The method of claim 3, further comprising, providing, by the processing unit, at least one control signal to a tensor traversal unit (TTU) to cause an array reference of the TTU to generate an address for a referenced array element used in executing the loop nest during performance of the tensor computation.

6. The method of claim 5, wherein the single instruction indicates a first TTU counter that is summed with a second TTU counter to generate an address for an array reference associated with the TTU.

7. The method of claim 1, wherein performing the tensor computation comprises, executing, by the processing unit, a first sync procedure that manages one or more operands associated with performance of the tensor computation, wherein managing an operand comprises stalling one or more loop nests based on a sync flag condition.

8. The method of claim 7, wherein performing the tensor computation comprises, executing, by the processing unit, a second sync procedure that manages incrementing a counter associated with a characteristic of the loop nest.

9. An electronic system for accelerating inference computations for a neural network having a plurality of neural network layers, the electronic system comprising:

one or more hardware compute units, each hardware compute unit including a processing unit and

one or more non-transitory machine-readable storage devices for storing instructions that are executable by the processing unit to cause performance of operations comprising:

receiving, by the processing unit of the hardware compute unit, a single instruction that specifies data values that:

encode a neural network layer type of a neural network layer of the plurality of neural network layers; and

are used by the hardware compute unit to perform a tensor computation for a neural network layer;

determining, by the processing unit and based on the single instruction, dimensions of at least one multi-dimensional tensor, where the at least one multi-dimensional tensor includes elements that correspond to an operand for performing the tensor computation;

generating, by the hardware compute unit and, based on the single instruction, a loop nest comprising a plurality of loops, wherein a structure of the loop nest is defined using the neural network layer type indicated by the data values of the single instruction; and

performing, by the hardware compute unit, the tensor computation by executing the loop nest to access operands corresponding to elements of the at least one multi-dimensional tensor.

10. The electronic system of claim 9, wherein the tensor computation is at least a portion of a computation of a neural network layer.

11. The electronic system of claim 9, wherein the single instruction causes the processing unit to access at least one memory address of an array in a storage medium, the memory address of the array comprising a variable that is read by the processing unit during performance of the tensor computation.

12. The electronic system of claim 11, wherein performing the tensor computation comprises, providing, by the processing unit, at least one control signal to a tensor traversal unit (TTU) to cause the TTU to emit loop indices used in executing the loop nest during performance of the tensor computation.

13. A non-transitory computer-readable storage medium comprising instructions executable by a processing unit which, upon such execution, causes the processing unit to perform the method of any of claims 1 to 8.