

SUPERPIXEL METHODS FOR CONVOLUTIONAL NEURAL NETWORKS

BACKGROUND

[0001] This specification relates to processing inputs through the layers of a neural network to generate outputs.

[0002] Neural networks are machine learning models that employ one or more layers of nonlinear units to predict an output 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 another layer in the network, e.g., the next hidden layer or the output layer. Each layer of the network generates an output from a received input in accordance with current values of a respective set of parameters.

SUMMARY

[0002a] According to an aspect of the present invention, there is provided a computer-implemented image recognition method for performing a computation of a convolutional neural network layer that is configured to receive an X by Y by Z input tensor representing an image and process the received input tensor using one or more convolutional neural network layer weight matrices to generate an U by V by W output tensor, the method comprising:

receiving the X by Y by Z input tensor that includes XY inputs, each of depth being Z, wherein X and Y represent spatial extent or index and Z represents depth extent or index,

transforming the X by Y by Z input tensor into a X' by Y' by Z' input tensor by grouping two or more of the XY inputs into X'Y' inputs while trading the spatial extent or index for the depth extent or index, wherein X' is smaller than or equal to X, Y' is smaller than or equal to Y, and Z' is larger than Z;

obtaining one or more modified convolutional neural network layer weight matrices, wherein the modified weight matrices are adapted to operate on the X' by Y' by Z' input tensor to generate a U' by V' by W' output tensor, wherein the U' by V' by W' output tensor includes U'V' outputs, each of depth being W', U' and V' represent spatial extent or index and W' represents depth extent or index;

processing the X' by Y' by Z' input tensor using the modified weight matrices to generate the U' by V' by W' output tensor; and

transforming the U' by V' by W' output tensor into a U by V by W output tensor by grouping two or more of the $U'V'$ outputs into UV outputs while trading the spatial extent or index for the depth extent or index, wherein U' is smaller than or equal to U , V' is smaller than or equal to V , and W' is larger than W ,

wherein the one or more convolutional neural network layer weight matrices comprise a number of kernel elements that operate over the XY inputs, and the obtaining step comprises:

generating the one or more modified convolutional neural network layer weight matrices by reducing the number of kernel elements to generate a reduced number of kernel elements that operate over the $X'Y'$ inputs.

[0003] In general, the subject matter described in this specification can be embodied in methods for more efficiently performing a computation of a convolutional neural network layer that is configured to receive an X by Y by Z input tensor and process the received input tensor using one or more convolutional neural network layer weight matrices to generate an U by V by W output tensor. The method includes transforming the X by Y by Z input tensor into a X' by Y' by Z' input tensor, wherein X' is smaller than or equal to X , Y' is smaller than or equal to Y , and Z' is larger than or equal to Z ; obtaining one or more modified convolutional neural network layer weight matrices, wherein (i) the modified weight matrices operate on the X' by Y' by Z' input tensor to generate a U' by V' by W' output tensor, and (ii) the U' by V' by W' output tensor comprises a transformed U by V by W output tensor, wherein U' is smaller than or equal to U , V' is smaller than or equal to V , and W' is larger than or equal to W ; and processing the X' by Y' by Z' input tensor using the modified weight matrices to generate the U' by V' by W' output tensor, wherein the U' by V' by W' output tensor comprises the U by V by W output tensor.

[0004] Other embodiments include corresponding computer systems, apparatus, and computer programs recorded on one or more computer storage devices, each configured to perform the actions of the methods. A system of one or more computers can be configured to perform particular operations or actions by virtue of software, firmware, hardware, or any combination thereof installed on the system that in operation may cause the system to perform the actions. One or more computer programs

WHAT IS CLAIMED IS:

1. A computer-implemented image recognition method for performing a computation of a convolutional neural network layer that is configured to receive an X by Y by Z input tensor representing an image and process the received input tensor using one or more convolutional neural network layer weight matrices to generate an U by V by W output tensor, the method comprising:

receiving the X by Y by Z input tensor that includes XY inputs, each of depth being Z, wherein X and Y represent spatial extent or index and Z represents depth extent or index,

transforming the X by Y by Z input tensor into a X' by Y' by Z' input tensor by grouping two or more of the XY inputs into X'Y' inputs while trading the spatial extent or index for the depth extent or index, wherein X' is smaller than or equal to X, Y' is smaller than or equal to Y, and Z' is larger than Z;

obtaining one or more modified convolutional neural network layer weight matrices, wherein the modified weight matrices are adapted to operate on the X' by Y' by Z' input tensor to generate a U' by V' by W' output tensor, wherein the U' by V' by W' output tensor includes U'V' outputs, each of depth being W', U' and V' represent spatial extent or index and W' represents depth extent or index;

processing the X' by Y' by Z' input tensor using the modified weight matrices to generate the U' by V' by W' output tensor; and

transforming the U' by V' by W' output tensor into a U by V by W output tensor by grouping two or more of the U'V' outputs into UV outputs while trading the spatial extent or index for the depth extent or index, wherein U' is smaller than or equal to U, V' is smaller than or equal to V, and W' is larger than W,

wherein the one or more convolutional neural network layer weight matrices comprise a number of kernel elements that operate over the XY inputs, and the obtaining step comprises:

generating the one or more modified convolutional neural network layer weight matrices by reducing the number of kernel elements to generate a reduced number of kernel elements that operate over the X'Y' inputs.

2. The method of claim 1, wherein X' is equal to the ceiling of X divided by a natural number N, Y' is equal to the ceiling of Y divided by a natural number M, and Z' is equal to Z multiplied by NM, and

preferably, the number of the plurality of first inputs equals NM inputs.

3. The method of claim 1, wherein the operation of the modified weight matrices on the X' by Y' by Z' input tensor is equivalent to the operation of the convolutional neural network layer weight matrices on the X by Y by Z input tensor.

4. The method of claim 1, wherein entries of the modified weight matrices comprise one or more convolutional neural network layer weight matrices.

5. The method of claim 1, wherein the modified weight matrices comprise matrices with a banded structure.

6. The method of claim 1, wherein U' is equal to the ceiling of U divided by a natural number P, V' is equal to the ceiling of V divided by a natural number Q, and W' is equal to W multiplied by PQ, and

preferably, the plurality of first outputs equals PQ outputs.

7. The method of claim 6, wherein the convolutional neural network layer includes a stride S in the X dimension and stride T in the Y dimension, and wherein the relationship between stride S, T, output tensor U by V by W and transformed output tensor U' by V' by W' is given by $U' = \text{ceiling}(U/NS)$ and $V' = \text{ceiling}(V/MT)$.

8. The method of claim 1, wherein the convolutional neural network layer comprises a pooling sub layer, and the method further comprises replacing the pooling sub layer with striding on the transformed input tensor.

9. The method of claim 1, wherein the convolutional neural network layer comprises kernel striding, and wherein the method further comprises implementing efficient kernel striding based on the transformed input tensor and modified weight matrices.

10. The method of claim 1, wherein the U by V by W output tensor represents UV first outputs and the U' by V' by W' output tensor represents U'V' super outputs with each super output comprising a plurality of first outputs, and

wherein the number of first outputs in the plurality of first outputs is dependent on the dimensions of the convolutional neural network layer weight matrices.

11. The method of claim 1, wherein the U by V by W output tensor represents UV first outputs and the U' by V' by W' output tensor represents U'V' super outputs with each super output comprising a plurality of first outputs, and

wherein the number of outputs in the plurality of first outputs is dependent on one or more of (i) the architecture of the convolutional neural network layer, (ii) an architecture of a neural network model including the convolutional neural network layer, or (iii) a device implementing the convolutional neural network layer.

12. The method of claim 1, wherein the modification of the one or more convolutional neural network layer weight matrices is a software only modification.

13. A system comprising one or more computers and one or more storage devices storing instructions that are operable, when executed by the one or more computers, to cause the one or more computers to perform the method according to any one of claims 1 to 12.

14. A computer storage medium encoded with instructions that, when executed by one or more computers, cause the one or more computers to perform the method according to any one of claims 1 to 12.