

CLAIMS

1. An information processing method, applied in a neural network and comprising:

inputting received input information into the neural network;

processing the input information by means of the neural network, wherein in the case that convolution processing is executed by means of a convolution layer of the neural network, a convolution kernel of the convolution layer is updated by using a transformation matrix configured for the convolution layer, so that the convolution processing of the convolution layer is completed by means of the updated convolution kernel; and

outputting a processing result of the processing of the neural network.

2. The method according to claim 1, wherein updating the convolution kernel of the convolution layer by using the transformation matrix configured for the convolution layer comprises:

acquiring a space dimension of the convolution kernel of the convolution layer;

executing duplication processing on the transformation matrix corresponding to the convolution layer based on the space dimension of the convolution kernel, wherein the number of times of duplication processing is determined by the space dimension of the convolution kernel; and

executing dot product processing on the transformation matrix after the duplication processing and the convolution kernel to obtain the updated convolution kernel of the corresponding convolution layer.

3. The method according to claim 1 or 2, before executing convolution processing by means of the convolution layer of the neural network, further comprising:

determining a matrix unit constituting the transformation matrix corresponding to the convolution layer, wherein the matrix unit comprises a first matrix and a second matrix, or only comprises the second matrix, wherein in response to the number of channels of an input feature and the number of channels of an output feature of the convolution layer being different, the transformation matrix corresponding to the convolution layer comprises the first matrix and the second matrix; and in response to the number of channels of the input

feature and the number of channels of the output feature of the convolution layer being identical, the transformation matrix corresponding to the convolution layer comprises the second matrix, wherein the first matrix is formed by connecting unit matrixes, and the second matrix is obtained by inner products of function transformations of a plurality of sub-matrixes; and

forming the transformation matrix of the convolution layer based on the determined matrix unit.

4. The method according to claim 3, wherein determining the second matrix constituting the transformation matrix of the convolution layer comprises:

acquiring a gate control parameter configured for each convolution layer;

determining the sub-matrixes constituting the second matrix based on the gate control parameter; and

forming the second matrix based on the determined sub-matrixes.

5. The method according to claim 4, wherein acquiring the gate control parameter configured for each convolution layer comprises:

acquiring the gate control parameter configured for each convolution layer according to received configuration information; or

determining the gate control parameter configured for the convolution layer based on a training result of the neural network.

6. The method according to any one of claims 3 to 5, wherein forming the transformation matrix of the convolution layer based on the determined matrix unit comprises:

acquiring the number of first channels of the input feature and the number of second channels of the output feature of each convolution layer;

in response to the number of the first channels being greater than the number of the second channels, forming the transformation matrix as a product of the first matrix and the second matrix; and

in response to the number of the first channels being less than the number of the second channels, forming the transformation matrix as a product of the second matrix and the first matrix.

7. The method according to claim 4 or 5, wherein determining the sub-matrixes constituting the second matrix based on the gate control parameter comprises:

performing function processing on the gate control parameter by using a sign function to obtain a binaryzation vector; and

obtaining a binaryzation gate control vector based on the binaryzation vector, and obtaining the plurality of sub-matrixes based on the binaryzation gate control vector, a first basic matrix, and a second basic matrix.

8. The method according to claim 7, wherein obtaining the binaryzation gate control vector based on the binaryzation vector comprises:

determining the binaryzation vector as the binaryzation gate control vector; or

determining a product of a permutation matrix and the binaryzation vector as the binaryzation gate control vector.

9. The method according to claim 7 or 8, wherein obtaining the plurality of sub-matrixes based on the binaryzation gate control vector, the first basic matrix, and the second basic matrix comprises:

in response to an element in the binaryzation gate control vector being a first numerical value, obtaining a sub-matrix of an all-ones matrix; and

in response to the element in the binaryzation gate control vector being a second numerical value, obtaining a sub-matrix of the unit matrix.

10. The method according to any one of claims 7 to 9, wherein the first basic matrix is the all-ones matrix, and the second basic matrix is the unit matrix.

11. The method according to any one of claims 4 to 10, wherein forming the second matrix based on the determined sub-matrixes comprises:

performing an inner product operation on the plurality of sub-matrixes to obtain the second matrix.

12. The method according to any one of claims 1 to 11, wherein the input information comprises at least one of text information, image information, video information, or voice information.

13. The method according to any one of claims 1 to 12, wherein the dimension of the transformation matrix is a product of the number of the first channels and the number of the second channels, the number of the first channels is the number of channels of the input feature of the convolution layer, the number of the second channels is the number of channels of the output feature of the convolution layer, and an element of the transformation matrix comprises at least one of 0 or 1.

14. The method according to any one of claims 1 to 13, further comprising a step of training the neural network, which comprises:

acquiring a training sample and a real detection result for monitoring;

performing processing on the training sample by using the neural network to obtain a prediction result; and

feeding back and adjusting a network parameter of the neural network based on a loss corresponding to the prediction result and the real detection result, until a termination condition is satisfied, wherein the network parameter comprises the convolution kernel of each network layer and the transformation matrix.

15. An information processing apparatus, comprising:

an input module configured to input received input information to a neural network;

an information processing module configured to process the input information by means of the neural network, wherein in the case that convolution processing is executed by means of a convolution layer of the neural network, a convolution kernel of the convolution layer is updated by using a transformation matrix configured for the convolution layer, so that the convolution processing of the convolution layer is completed by means of the updated convolution kernel; and

an output module configured to output a processing result of the processing of the neural network.

16. The apparatus according to claim 15, wherein the information processing module is further configured to: acquire a space dimension of the convolution kernel of the convolution layer;

execute duplication processing on the transformation matrix corresponding to the convolution layer based on the space dimension of the convolution kernel, wherein the number of times of duplication processing is determined by the space dimension of the convolution kernel; and

execute dot product processing on the transformation matrix after the duplication processing and the convolution kernel to obtain the updated convolution kernel of the corresponding convolution layer.

17. The apparatus according to claim 15 or 16, wherein the information processing module is further configured to determine a matrix unit constituting the transformation matrix corresponding to the convolution layer, and form the transformation matrix of the convolution layer based on the determined matrix unit, wherein the matrix unit comprises a first matrix and a second matrix, or only comprises the second matrix, wherein in response to the number of channels of an input feature and the number of channels of an output feature of the convolution layer being different, the transformation matrix corresponding to the convolution layer comprises the first matrix and the second matrix; and in response to the number of channels of the input feature and the number of channels of the output feature of the convolution layer being identical, the transformation matrix corresponding to the convolution layer comprises the second matrix, wherein the first matrix is formed by connecting unit matrixes, and the second matrix is obtained by inner products of function transformations of a plurality of sub-matrixes.

18. An electronic device, comprising:

a processor; and

a memory configured to store processor-executable instructions;

wherein the processor is configured to call the instructions stored in the memory, so as to execute the method according to any one of claims 1 to 14.

19. A computer-readable storage medium, having computer program instructions stored thereon, wherein when the computer program instructions are executed by a processor, the method according to any one of claims 1 to 14 is implemented.

20. A computer program, comprising a computer readable code, wherein when the

computer readable code runs in an electronic device, a processor in the electronic device executes instructions for implementing the method according to any one of claims 1 to 14.