## CLAIMS

What is claimed is

1. A task processing method based on a neural network, comprising:

acquiring at least two network modules by dividing network layers in the neural network based on a number of channels between two adjacent network layers in the neural network, wherein one network module comprises at least one network layer;

acquiring input data, wherein the input data is intended to trigger thread tasks, and is source input data or cache exchange data;

generating processing result data by scheduling at least two corresponding module threads in parallel based on at least two triggered thread tasks to process the input data, wherein the at least two module threads respectively correspond to the at least two network modules based on division of the network layers in the neural network; and

outputting the processing result data to a cache as the cache exchange data of module threads except the at least two module threads, or outputting the processing result data as a processing result of the source input data.

~~2. The method according to claim 1, wherein prior to acquiring the input data, the method further comprises:~~

~~acquiring the at least two network modules by dividing the network layers in the neural network.~~

~~3~~2. The method according to claim ~~2~~1, wherein acquiring the at least two network modules by dividing the network layers in the neural network based on the number of channels between the two adjacent network layers in the neural network~~acquiring the at least two network modules by dividing the network layer in the neural network~~ comprises:

determining ~~a~~the number of channels between adjacent network layers in the neural network;

dividing a former network layer in the adjacent network layers as an input layer for inputting the network module and dividing a latter network layer in the adjacent network layers as an output layer for outputting the network module in response to determining that one channel is present between the adjacent network layers; and

generating the at least two network modules based on the output layer for outputting the network module and the input layer for inputting the network module.

43. The method according to claim 1, wherein a difference value between processing times of the network modules is less than a predetermined threshold.

54. The method according to claim 1, wherein the at least two module threads at least comprise a start module thread and an end module thread;

scheduling the start module thread based on the triggered thread task to process the input data comprises: calling the start module thread to pre-process the source input data, performing task processing based on a start network module of the neural network corresponding to the start module thread, and outputting the processing result data to the cache as the cache exchange data; and

scheduling the end module thread based on the triggered thread task to process the input data comprises: calling the end module thread to acquire the corresponding cache exchange data from the cache as the input data, performing task processing based on an end network module of the neural network corresponding to the end module thread, and post-processing and outputting processing result data as the processing result of the source input data.

65. The method according to claim 54, wherein the module threads further comprise at least one intermediate module thread; and scheduling the at least one intermediate module thread based on the triggered thread task to process the input data comprises:

calling the at least one intermediate module thread to acquire the corresponding cache exchange data from the cache as the input data, performing task processing based on an intermediate network module of the neural network corresponding to the intermediate module thread, and outputting the processing result data to the cache as the cache exchange data.

76. The method according to claim 5 4 or 65, wherein the source input data is image frame data, the pre-processing is image pre-processing, the post-processing is image post-processing, and the neural network is a convolutional neural network.

87. The method according to claim 1, wherein acquiring the input data comprises:

determining a triggering condition for generating the thread tasks in response to monitoring that data is written in a target cache space correspondingly read by the thread tasks, and determining the data written in the target cache space as the input data, wherein the module threads are bound to the thread tasks in one-to-one correspondence, and the thread tasks are bound to the target cache space.

~~9~~8. The method according to claim 1, wherein the module threads are bound to processor cores in one-to-one correspondence.

~~10~~9. A task processing apparatus based on a neural network, comprising:

an input data acquiring module, configured to acquire input data, wherein the input data is intended to trigger thread tasks, and is source input data or cache exchange data;

a module thread scheduling module, configured to generate processing result data by scheduling at least two corresponding module threads in parallel based on at least two triggered thread tasks to process the input data, wherein the at least two module threads respectively correspond to at least two network modules based on division of network layers in the neural network; and

a processing result data outputting module, configured to output the processing result data to a cache as the cache exchange data of module threads except the at least two module threads, or output the processing result data as a processing result of the source input data;

wherein the task processing apparatus based on the neural network is further configured to acquire the at least two network modules by dividing the network layers in the neural network based on a number of channels between two adjacent network layers in the neural network, wherein one network module comprises at least one network layer.

~~11~~10. A task processing device based on a neural network, comprising: a processor and a memory storing at least one instruction therein;

wherein the instruction, when executed by the processor, causes the device to execute the task processing method based on the neural network as defined in any one of claims 1 to ~~9~~8.

~~12~~11. A computer-readable storage medium, storing at least one instruction therein, wherein the at least one instruction, when executed by a processor of a device, causes the device to execute the task processing method based on the neural network as defined in any one of claims 1 to ~~9~~8.