

to determine an output data corresponding to each of the multiple pieces of input data; determining a loss value based on the output data and a label corresponding to the input data; and determining the present sampling network as a standby network when the loss value is less than or equal to a preset threshold.

[9] In some possible embodiments, the method further includes: before determining the present sampling network as the standby network when the loss value is less than or equal to the preset threshold, back propagating based on the loss value when the loss value is greater than the preset threshold, performing parameter update on the present sampling network to obtain an updated sampling network, and determining the updated sampling network as the present sampling network; repeating the following: processing the multiple pieces of input data based on the present sampling network to determine the output data corresponding to each of the multiple pieces of input data, and determining the loss value based on the output data and the label corresponding to the input data.

[10] In some possible embodiments, training the multiple sampling networks with the sample data set and assigning the multiple sampling networks trained as the multiple standby networks includes: assigning the first level network as a present first sampling network, and assigning each of at least one second level work and the third level network as a present second sampling network; processing the multiple pieces of input data based on the present first sampling network to determine a first output data corresponding to each input data, and determining a first loss value based on the first output data corresponding to each input data and a label corresponding to each input data; for each of multiple present second sampling networks: processing the multiple pieces of input data based on the present second sampling network to determine a second output data corresponding to each input data, and determining a second loss value based on the second output data corresponding to each input data and the first output data corresponding to each input data; determining the present first sampling network and the present second sampling network as the standby network, when the first loss value is less than or equal to a first preset threshold and the second loss value is less than or equal to a second preset threshold.

[11] In some possible embodiments, when the first loss value is less than or equal to the first preset threshold and the second loss value is less than or equal to the second preset threshold, the method further includes: before determining the present first sampling network and the present second sampling network as the standby network, back propagating based on the first loss value when the first loss value is greater than the first preset threshold, performing parameter update on the present first sampling network to obtain

layer, and the activation layer are connected in series. The convolutional layer is a depthwise convolutional layer, a pointwise convolutional layer, or a two-dimensional convolutional layer.

[26] In some possible embodiments, the device resource information of the original device at least includes: video-memory capacity data, video-memory performance data, network performance data, memory capacity data, and computing ability data of the original device, the device resource information of the target device at least includes: video-memory capacity data, video-memory performance data, network performance data, memory capacity data, and computing ability data of the target device.

[27] In some possible embodiments, the original network determining unit is configured to: decompose the target event to determine a target object identifier and a target task identifier of the target event; determine a task identifier and an object identifier of each network in the network set; determine a first match degree value between the target task identifier and each task identifier; determine a second match degree value between the target object identifier and each object identifier; determine the original network from the network set according to the first match degree value and the second match degree value.

[28] In some possible embodiments, the target network dividing unit is configured to: adjust an output resolution, the number of output channels, and the number of convolutional modules of each network module in the original network according to the device resource information of the target device and the device resource information of the original device, with a preset search rule; and determine the original network adjusted as the original network.

[29] In some possible embodiments, an electronic device is provided. The electronic device includes at least one processor and a memory in communication with the at least one processor. The memory stores instructions which are executable with the at least one processor, the processor is configured to execute the instructions stored in the memory to perform the method for determining a deep learning network of the disclosure.

[30] In some possible embodiments, a computer readable storage medium is provided. The computer storage medium stores at least one instruction or at least one program, the at least one instruction or the at least one program are uploaded and executed by a processor to perform the method for determining a deep learning network of the disclosure.

[31] In some possible embodiments, a computer program product including instructions is provided. When the instructions are running on a computer, are operable with the computer to perform the method for determining a deep learning network of the disclosure.

[32] In embodiments of the disclosure, a target event to be processed by a target device

the manner in which the object is processed is “segmentation”, therefore, the task ID of “image segmentation” can be the ID corresponding to “segmentation”, and the object ID can be the ID corresponding to “image”. Still another example, “article scoring” event can be decomposed into: an object to be processed is an “article”, and the manner in which the object is processed is “scoring”, therefore, the task ID of “article scoring” can be the ID corresponding to “scoring”, and the object ID can be the ID corresponding to “article”. In some possible embodiments, the task ID and the object ID are unique IDs, for example, in the object ID, the ID corresponding to “image” is A001, the ID corresponding to “Article” is A002....., in the task ID, the ID corresponding to “recognition” is B001, the ID corresponding to “segmentation” is B002, and the ID corresponding to “scoring” is B003, etc.

[72] Based on the above examples, at S2021, if the target event is “image segmentation”, the server can determine that the target task ID of the target event is B002 and the target object ID of the target event is A001.

[73] At S2022, the task ID and object ID for each network in the network set is obtained.

[74] In some possible embodiments, when the server obtaining various networks, each network may carry its own task ID and object ID. In some other possible embodiments, when the server obtaining various networks, each network may carry the ID of an event, and the server can decompose the ID of the event to obtain the task ID and the object ID of each event.

[75] At S2023, a first match degree value between the target task ID and each task ID is determined.

[76] At S2024, a second match degree value between the target object ID and each object ID is determined.

[77] To be continued, the target event is image segmentation. In some possible embodiments, the target task ID is compared with the task ID of each network in the network set, if they are the same, then the first match degree value is 1, otherwise, the first match degree value is 0. Similarly, the target object ID is compared with the object ID of each network in the network set, if they are the same, then the second match degree value is 1, otherwise, the second match degree value is 0.

[78] However, the above matching manner is too absolute. For example, if the target object ID is “image A001” and the object ID is “picture A101”, according to the above method, the second match degree value is 0 and this is not reasonable and may even lead to the failure of the whole scheme. Therefore, in some possible embodiments, take the target

object ID as an example, similar extended object ID can be determined according to the target object ID, for example, similar extended object ID “picture A101”, “photo A012” and so on can be determined according to “image A001”. Thereafter, the target object ID and the extended object ID can each be compared with the object ID to obtain the second match degree value between the target object ID and the object ID. Specifically, the manner in which the first match degree value between the target task ID and the task ID is determined can make reference to that of the object ID, extended task IDs corresponding to the target task ID can be determined first.

[79] At S2025, the original network is determined from the network set according to the first match degree value and the second match degree value.

[80] In some possible embodiments, a network with both the first match degree value and the second match degree value of 1 can be determined as the original network.

[81] Embodiments of the disclosure further includes a procedure of training to obtain the original network. In one embodiment, the original network includes but is not limited to deep learning networks such as a convolutional neural network, a recurrent neural network or a recursive neural network. Deep learning network can be considered as a form of machine learning (ML). ML is a multi-domain interdisciplinary subject, involving probability theory, statistics, approximation theory, convex analysis, algorithm complexity theory and so on. ML focuses on how computers simulate or realize human learning behavior, so as to obtain new knowledge or skills, reorganize the existing knowledge structure and continuously improve its performance. ML is the core of artificial intelligence and the fundamental way to make computers intelligent, it is applied in all fields of artificial intelligence. ML and deep learning usually include artificial neural network, confidence network, reinforcement learning, transfer learning, inductive learning, type teaching learning and so on. ML can be divided into supervised machine learning, unsupervised machine learning and semi supervised machine learning.

[82] In one embodiment, taking the convolution neural network as an example, FIG. 4 is a flowchart illustrating a procedure of training to obtain the original network. As illustrated in FIG. 4, the method includes the following.

[83] At S401, a sample data set is obtained, the sample data set includes multiple pieces of input data and labels each corresponding to one piece of input data.

[84] If an image segmentation event is previously processed by the original network, before training the original network, two preparations need to be made: i) construct a preset learning network; ii) collecting a sample data set. Take image segmentation as an example,

device resource information of the original device. For example, the original device is a desktop and the target device is a mobile phone, for each network module, the output resolution, the number of output channels, and the number of convolutional modules can be reduced. Or, the device resource information of the target device is basically similar to the device resource information of the original device except that there is one device resource of the target device greater than the same device resource of the original device, therefore, for some network modules, the output resolution, the number of output channels, and the number of convolutional modules can be increased.

[122] In some possible embodiments, to be on the safe side and making the target network more suitable for the target device, after the target network is deployed on the target device, the target network can run on the target device, so as to determine whether further adjustment on the output resolution, the number of output channels, and the number of convolutional modules are required based on the running result, so as to make sure the target network is more suitable for the target device.

[123] In some possible embodiments, in order to decouple training and search and obtain better performance of the target network, the process of training the original network and searching for the target network from the original network can be separated as described above. Alternatively, the process of training the original network and the process of searching for the target network from the original network can be intercross carried out. This can be determined according to actual needs and is not limited herein.

[124] To be summarized, embodiments of the disclosure provide a preset learning network (Hyper network, such as the two-branch network). After training, the target network suitable for various target devices can be obtained from the original network with a network structure search algorithm according to the device resource information obtained without training each target network, this can not only reduce the time spent by researchers in network design and training, but also make the trained original network more valuable.

[125] FIG. 9 is a module diagram illustrating an apparatus for determining a deep learning network. As illustrated in FIG. 9, the apparatus 900 for determining a deep learning network includes a target event determining unit 901, an original network determining unit 902, a resource information obtaining unit 903, and a target network dividing unit 904.

[126] The target event determining unit 901 is configured to determine a target event to be processed by a target device. The original network determining unit 902 is configured to determine an original network from a trained network set according to the target event, the original network being a network in the network set which matches the target event. The

What is claimed is:

1. A method for determining a deep learning network, comprising:
 - determining a target event to be processed by a target device;
 - determining an original network from a trained network set according to the target event, the original network being a network in the network set which matches the target event;
 - obtaining device resource information of an original device that trains the original network and device resource information of the target device; and
 - dividing a target network from the original network according to the device resource information of the target device and the device resource information of the original device, the target network being to be deployed on the target device to process the target event, the target network and the original network being the same in the number of network modules and a connection structure between network modules.
2. The method of claim 1, further comprising training to obtain the original network:
 - obtaining a sample data set, the sample data set comprising a plurality of input data and labels each corresponding to one piece of input data;
 - constructing a preset learning network, determining a plurality of sampling networks from the preset learning network, and setting an initial parameter for the sampling network; and
 - training the plurality of sampling networks with the sample data set and assigning the plurality of sampling networks trained as a plurality of standby networks, and determining the original network from the plurality of standby networks.
3. The method of claim 2, wherein the plurality of sampling networks comprises a first level network, at least one second level network, and a third level network, the first level network, the at least one second level network, and the third level network are same in the number of network modules and a connection structure between network modules;
 - each network module in the first level network having an output resolution greater than a corresponding network module in the second level network, and each network module in the second level network having an output resolution greater than a corresponding network module in the third level network;
 - each network module in the first level network having an output channel quantity

greater than a corresponding network module in the second level network, and each network module in the second level network having an output channel quantity greater than a corresponding network module in the third level network;

each network module in the first level network having a convolution module quantity greater than a corresponding network module in the second level network, and each network module in the second level network having a convolution module quantity greater than a corresponding network module in the third level network.

4. The method of claim 2 or 3, wherein training the plurality sampling networks with the sample data set and assigning the plurality sampling networks trained as the plurality of standby networks comprises:

assigning each of the plurality of sampling networks as a present sampling network;
processing the plurality of input data based on the present sampling network to determine an output data corresponding to each of the plurality of input data;

determining a loss value based on the output data and a label corresponding to the input data; and

determining the present sampling network as a standby network when the loss value is less than or equal to a preset threshold.

5. The method of claim 4, further comprising:

before determining the present sampling network as the standby network when the loss value is less than or equal to the preset threshold,

back propagating based on the loss value when the loss value is greater than the preset threshold, performing parameter update on the present sampling network to obtain an updated sampling network, and determining the updated sampling network as the present sampling network; and

repeating the following:

processing the plurality of input data based on the present sampling network to determine the output data corresponding to each of the plurality of input data; and

determining the loss value based on the output data and the label corresponding to the input data.

6. The method of claim 3, wherein training the plurality sampling networks with the sample data set and assigning the plurality sampling networks trained as the plurality of standby networks comprises:

assigning the first level network as a present first sampling network, and assigning each of at least one second level work and the third level network as a present second sampling network;

processing the plurality of input data based on the present first sampling network to determine a first output data corresponding to each input data, and determining a first loss value based on the first output data corresponding to each input data and a label corresponding to each input data;

for each of a plurality of present second sampling networks: processing the plurality of input data based on the present second sampling network to determine a second output data corresponding to each input data, and determining a second loss value based on the second output data corresponding to each input data and the first output data corresponding to each input data;

determining the present first sampling network and the present second sampling network as the standby network, when the first loss value is less than or equal to a first preset threshold and the second loss value is less than or equal to a second preset threshold.

7. The method of claim 6, further comprising:

before determining the present first sampling network and the present second sampling network as the standby network when the first loss value is less than or equal to the first preset threshold and the second loss value is less than or equal to the second preset threshold,

back propagating based on the first loss value when the first loss value is greater than the first preset threshold, performing parameter update on the present first sampling network to obtain an updated first sampling network, and determining the updated first sampling network as the present first sampling network; back propagating based on the second loss value when the second loss value is greater than the second preset threshold, performing parameter update on the present second sampling network to obtain an updated second sampling network, and determining the updated second sampling network as the present second sampling network; and repeating the following:

processing the plurality of input data based on the present first sampling network to determine the first output data corresponding to each input data, and determining the first loss value based on the first output data corresponding to each input data and the label corresponding to each input data; and

processing the plurality of input data based on the present second sampling network to determine the second output data corresponding to each input data, and determining the second loss value based on the second output data corresponding to each input data and the first output data corresponding to each input data.

8. The method of claim 1 or 2, wherein the original network comprises one main branch and at least one slave branch, the main branch comprises a plurality of network modules, and network modules in each slave branch comprises part of the network modules in the main branch, wherein

if the original network has a plurality of slave branches, each of the plurality of branches shares at least one network module.

9. The method of claim 8, wherein each network module comprises at least one convolutional module, and each convolutional module comprises one convolutional layer, one batch normalization (BN) layer, and one activation layer, wherein

the convolutional layer, the BN layer, and the activation layer are connected in series;

the convolutional layer is a depthwise convolutional layer, a pointwise convolutional layer, or a two-dimensional convolutional layer.

10. The method of claim 1, wherein the device resource information of the original device at least comprises: video-memory capacity data, video-memory performance data, network performance data, memory capacity data, and computing ability data of the original device, the device resource information of the target device at least comprises: video-memory capacity data, video-memory performance data, network performance data, memory capacity data, and computing ability data of the target device.

11. The method of claim 1, wherein determining the original network from the trained

network set according to the target event comprises:

- decomposing the target event to determine a target object identifier and a target task identifier of the target event;

- obtaining a task identifier and an object identifier of each network in the network set;

- determining a first match degree value between the target task identifier and each task identifier;

- determining a second match degree value between the target object identifier and each object identifier; and

- determining the original network from the network set according to the first match degree value and the second match degree value.

12. The method of claim 1, wherein dividing the target network from the original network according to the device resource information of the target device and the device resource information of the original device comprises:

- adjusting an output resolution, an output channel quantity, and a convolutional module quantity of each network module in the original network according to the device resource information of the target device and the device resource information of the original device, with a preset search rule; and

- determining the original network adjusted as the original network.

13. An apparatus for determining a deep learning network, comprising:

- a target event determining unit configured to determine a target event to be processed by a target device;

- an original deep learning network determining unit configured to determine an original network from a trained network set according to the target event, the original network being a network in the network set which matches the target event;

- a resource information obtaining unit configured to obtain device resource information of an original device that trains the original network and device resource information of the target device;

- a target network dividing unit configured to divide a target network from the original network according to the device resource information of the target device and the device resource information of the original device, the target network being deployed on the target device to process the target event, the target network and the original network being the same

in the number of network modules and a connection structure between network modules.

14. An electronic device, comprising at least one processor and a memory in communication with the at least one processor, the memory storing instructions which are executable with the at least one processor, the processor being configured to execute the instructions stored in the memory to perform the method of any of claims 1 to 12.

15. A computer readable storage medium, storing at least one instruction or at least one program, the at least one instruction or the at least one program being uploaded and executed by a processor to perform the method of any of claims 1 to 12.