

NEURAL NETWORK CONSTRUCTION METHOD AND SYSTEM

[0001]

TECHNICAL FIELD

5 [0002] This application relates to the field of artificial intelligence, and in particular, to a neural network construction method and system.

BACKGROUND

10

[0003] In the field of artificial intelligence, a neural network has made outstanding achievements in processing and analyzing a plurality of multimedia signals such as an image, a video and a voice in recent years. A well-performing neural network usually has a delicate network architecture that requires a great deal of effort to be designed by highly skilled and rich experienced human experts. It is impractical for an ordinary non-expert user to design the neural network for specific problems. A backbone network is an important network architecture in the neural network, and is usually not associated with a specific task of the neural network. For example, the backbone network may be a feature extraction network or a residual network required for most tasks. For example, a common computer vision task is to perform object detection and semantic segmentation. 15 Usually, the backbone network may first perform feature extraction on input information, and then input an extracted feature to a prediction module, to obtain a prediction result. A more complex to-be-processed scenario calls for an improved feature extraction capability of the backbone network, and correspondingly design of the backbone network becomes more difficult.

[0004] Some solutions for architecture search of a model have been proposed. However, 25 architecture search in an existing solution is usually only for a single task, and a network architecture needs to be redesigned or re-searched in different scenarios. Therefore, a large amount

CLAIMS

1. A neural network construction method, comprising:
 - obtaining an initial backbone network and a candidate set, wherein the initial backbone network is used for constructing a target neural network;
 - 5 replacing at least one basic unit in the initial backbone network with at least one placeholder module to obtain a to-be-determined network, wherein the candidate set comprises parameters of a plurality of structures corresponding to the at least one placeholder module;
 - performing sampling based on the candidate set to obtain information about at least one sampling structure;
 - 10 obtaining a network model based on the to-be-determined network and the information about the at least one sampling structure, wherein the information about the at least one sampling structure is used for determining a structure of the at least one placeholder module; and
 - if the network model meets a preset condition, using the network model as the target neural network.
- 15 2. The method according to claim 1, wherein after the obtaining a network model based on the to-be-determined network and the information about the at least one sampling structure, the method further comprises:
 - if the network model does not meet the preset condition, performing resampling based on the candidate set, and updating the network model based on the information that is about the at least
 - 20 one sampling structure and that is obtained through resampling.
3. The method according to claim 1, wherein before the performing sampling based on the candidate set to obtain information about at least one sampling structure, the method comprises:
 - constructing a parameter space based on the candidate set, wherein the parameter space comprises architecture parameters corresponding to the parameters of the plurality of structures;
 - 25 and
 - the performing sampling based on the candidate set to obtain information about at least one sampling structure comprises:
 - performing sampling on the parameter space to obtain at least one group of sampling parameters corresponding to the at least one sampling structure.
- 30 4. The method according to claim 3, wherein the obtaining a network model based on the to-

be-determined network and the information about the at least one sampling structure comprises:

converting the structure of the at least one placeholder module in the to-be-determined network based on the at least one group of sampling parameters, to obtain the network model.

5 5. The method according to claim 3, wherein before the obtaining a network model based on the to-be-determined network and the information about the at least one sampling structure, the method further comprises:

constructing the plurality of structures based on the candidate set and the to-be-determined network, wherein the plurality of structures form a structure search space; and

10 the obtaining a network model based on the to-be-determined network and the information about the at least one sampling structure comprises:

searching the network model from the structure search space based on the at least one group of sampling parameters.

6. The method according to claim 1, wherein the preset condition comprises one or more of the following: a quantity of times of obtaining the network model exceeds a preset quantity of times, duration for obtaining the network model exceeds preset duration, or an output result of the network model meets a preset requirement.

7. The method according to claim 1, wherein the candidate set comprises one or more of the following: a type of an operator, attribute information of an operator, or a connection mode between operators.

20 8. The method according to claim 1, wherein the target neural network is used for performing at least one of picture recognition, semantic segmentation, or object detection.

9. A neural network construction apparatus, comprising a processor, wherein the processor is coupled to a memory, the memory stores a program, and when program instructions stored in the memory are executed by the processor, the neural network construction apparatus is enabled to perform the method:

obtaining an initial backbone network and a candidate set, wherein the initial backbone network is used for constructing a target neural network;

replacing at least one basic unit in the initial backbone network with at least one placeholder module to obtain a to-be-determined network, wherein the candidate set comprises parameters of a plurality of structures corresponding to the at least one placeholder module;

performing sampling based on the candidate set to obtain information about at least one

sampling structure;

obtaining a network model based on the to-be-determined network and the information about the at least one sampling structure, wherein the information about the at least one sampling structure is used for determining a structure of the at least one placeholder module; and

5 if the network model meets a preset condition, using the network model as the target neural network.

10. The apparatus according to claim 9, wherein after the obtaining a network model based on the to-be-determined network and the information about the at least one sampling structure, the method further comprises:

10 if the network model does not meet the preset condition, performing resampling based on the candidate set, and updating the network model based on the information that is about the at least one sampling structure and that is obtained through resampling.

11. The apparatus according to claim 9, wherein before the performing sampling based on the candidate set to obtain information about at least one sampling structure, the method comprises:

15 constructing a parameter space based on the candidate set, wherein the parameter space comprises architecture parameters corresponding to the parameters of the plurality of structures; and

the performing sampling based on the candidate set to obtain information about at least one sampling structure comprises:

20 performing sampling on the parameter space to obtain at least one group of sampling parameters corresponding to the at least one sampling structure.

12. The apparatus according to claim 11, wherein the obtaining a network model based on the to-be-determined network and the information about the at least one sampling structure comprises:

25 converting the structure of the at least one placeholder module in the to-be-determined network based on the at least one group of sampling parameters, to obtain the network model.

13. The apparatus according to claim 12, wherein before the obtaining a network model based on the to-be-determined network and the information about the at least one sampling structure, the method further comprises:

30 constructing the plurality of structures based on the candidate set and the to-be-determined network, wherein the plurality of structures form a structure search space; and

the obtaining a network model based on the to-be-determined network and the information

about the at least one sampling structure comprises:

searching the network model from the structure search space based on the at least one group of sampling parameters.

14. The apparatus according to claim 9, wherein the preset condition comprises one or more of the following: a quantity of times of obtaining the network model exceeds a preset quantity of times, duration for obtaining the network model exceeds preset duration, or an output result of the network model meets a preset requirement.

15. The apparatus according to claim 9, wherein the candidate set comprises one or more of the following: a type of an operator, attribute information of an operator, or a connection mode between operators.

16. The apparatus according to claim 9, wherein the target neural network is used for performing at least one of picture recognition, semantic segmentation, or object detection.

17. A computer-readable storage medium, wherein the computer-readable storage medium stores a computer program, when the computer program runs on a computer, the computer is enabled to perform the device enabling method:

obtaining an initial backbone network and a candidate set, wherein the initial backbone network is used for constructing a target neural network;

replacing at least one basic unit in the initial backbone network with at least one placeholder module to obtain a to-be-determined network, wherein the candidate set comprises parameters of a plurality of structures corresponding to the at least one placeholder module;

performing sampling based on the candidate set to obtain information about at least one sampling structure;

obtaining a network model based on the to-be-determined network and the information about the at least one sampling structure, wherein the information about the at least one sampling structure is used for determining a structure of the at least one placeholder module; and

if the network model meets a preset condition, using the network model as the target neural network.

18. The computer-readable storage medium according to claim 17, wherein after the obtaining a network model based on the to-be-determined network and the information about the at least one sampling structure, the method further comprises:

if the network model does not meet the preset condition, performing resampling based on the

candidate set, and updating the network model based on the information that is about the at least one sampling structure and that is obtained through resampling.

19. The computer-readable storage medium according to claim 17, wherein before the performing sampling based on the candidate set to obtain information about at least one sampling
5 structure, the method comprises:

constructing a parameter space based on the candidate set, wherein the parameter space comprises architecture parameters corresponding to the parameters of the plurality of structures;
and

the performing sampling based on the candidate set to obtain information about at least one
10 sampling structure comprises:

performing sampling on the parameter space to obtain at least one group of sampling parameters corresponding to the at least one sampling structure.

20. The computer-readable storage medium according to claim 19, wherein the obtaining a network model based on the to-be-determined network and the information about the at least one
15 sampling structure comprises:

converting the structure of the at least one placeholder module in the to-be-determined network based on the at least one group of sampling parameters, to obtain the network model.