

NEURAL NETWORK CONSTRUCTION METHOD AND SYSTEM

[0001] ~~This application claims priority to Chinese Patent Application No. 202010425173.3, filed with the China National Intellectual Property Administration on May 19, 2020 and entitled "NEURAL NETWORK CONSTRUCTION METHOD AND SYSTEM", which is incorporated herein by reference in its entirety.~~

TECHNICAL FIELD

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

BACKGROUND

[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. 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, 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

What is claimed is:

1. A neural network construction method, comprising:

obtaining an initial backbone network and a candidate set, wherein the initial backbone
5 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
10 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
15 network.

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
20 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.

3. The method according to claim 1 or 2, 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
25 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:

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

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. 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

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 any one of claims 1 to 5, wherein a sampling mode of the performing sampling based on the candidate set comprises: random sampling or sampling according to a preset rule.

7. The method according to claim 6, wherein if the sampling mode is sampling according to the preset rule, after it is determined that the network model does not meet the preset condition, the method further comprises:

updating the preset rule by using a preset optimization algorithm based on an estimation result of the network model.

8. The method according to claim 7, wherein the optimization algorithm comprises an evolutionary algorithm, a reinforcement learning algorithm, a Bayesian optimization algorithm, or a gradient optimization algorithm.

9. The method according to any one of claims 1 to 8, 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.

10. The method according to any one of claims 1 to 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.

11. The method according to any one of claims 1 to 10, wherein the target neural network is used for performing at least one of picture recognition, semantic segmentation, or object detection.

12. The method according to any one of claims 1 to 11, wherein after the using the network model as the target neural network, the method further comprises:

5 training the target neural network based on a preset data set, to obtain the trained target neural network.

13. The method according to any one of claims 1 to 12, wherein the obtaining an initial backbone network and a candidate set comprises:

obtaining the initial backbone network and the candidate set from user input data.

10 14. A neural network construction system, comprising an input module, a sampling module, an architecture constructor, and an architecture estimator, wherein

the input module is configured to obtain an initial backbone network and a candidate set, wherein the initial backbone network is used for constructing a target neural network;

15 the architecture constructor is configured to replace 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;

the sampling module is configured to perform sampling based on the candidate set to obtain information about at least one sampling structure;

20 the architecture constructor is further configured to obtain 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

25 the architecture estimator is configured to estimate whether the network model meets a preset condition, and if the network model meets the preset condition, use the network model as the target neural network.

15. The neural network construction system according to claim 14, wherein

30 if the network model does not meet the preset condition, the sampling module is further configured to perform resampling based on the candidate set, and the architecture constructor is further configured to update the network model based on the information that is about the at least one sampling structure and that is obtained through resampling.

16. The neural network construction system according to claim 14 or 15, wherein the architecture constructor is further configured to construct 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

5 the sampling module is specifically configured to perform sampling on the parameter space to obtain at least one group of sampling parameters corresponding to the at least one sampling structure.

17. The neural network construction system according to claim 16, wherein the architecture constructor is specifically configured to convert the structure of the at least
10 one placeholder module in the to-be-determined network based on the at least one sampling structure, to obtain the network model.

18. The neural network construction system according to claim 16, wherein the architecture constructor is further configured to: before obtaining the network model based on the to-be-determined network and the information about the at least one sampling
15 structure, construct 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 architecture constructor is specifically configured to search the network model from the structure search space based on the at least one group of sampling parameters.

19. The neural network construction system according to any one of claims 14 to 18, wherein
20 a sampling mode of the performing sampling based on the candidate set comprises: random sampling or sampling according to a preset rule.

20. The neural network construction system according to claim 19, wherein if the sampling mode is sampling according to the preset rule, after it is determined that the network model does not meet the preset condition, the sampling module is further configured to update the preset rule
25 by using a preset optimization algorithm based on an estimation result of the network model.

21. The neural network construction system according to claim 20, wherein the optimization algorithm comprises an evolutionary algorithm, a reinforcement learning algorithm, a Bayesian optimization algorithm, or a gradient optimization algorithm.

22. The neural network construction system according to any one of claims 14 to 21, wherein
30 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.

~~23. The neural network construction system according to any one of claims 14 to 22, 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.~~

~~24. The neural network construction system according to any one of claims 14 to 23, wherein the target neural network is used for performing at least one of picture recognition, semantic segmentation, or object detection.~~

~~25. The neural network construction system according to any one of claims 14 to 24, wherein the neural network construction system further comprises:~~

~~a training module, configured to: after the using the network model as the target neural network, train the target neural network based on a preset data set, to obtain the trained target neural network.~~

~~26. The neural network construction system according to any one of claims 14 to 25, wherein the input module is specifically configured to:~~

~~obtain the initial backbone network and the candidate set from user input data.~~

~~27. 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 method according to any one of claims 1 to 13 is implemented.~~

~~28. A computer readable storage medium, comprising a program, wherein when the program is executed by a processing unit, the method according to any one of claims 1 to 13 is performed.~~

~~29. A neural network construction apparatus, comprising a processing unit and a communication interface, wherein the processing unit obtains program instructions through the communication interface, and when the program instructions are executed by the processing unit, the method according to any one of claims 1 to 13 is implemented.~~ 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;

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.

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 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; 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.

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. 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

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.

5 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.

10 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.

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.

15 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;

20 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;

25 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.

30 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:

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.

5 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:

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

10 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.

15 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:

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.

20 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:

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:

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

30 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.

5 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;

10 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;

15 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.

20 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.

25 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 structure, the method comprises:

30 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:

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

5 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 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.

10