

CLAIMS

1. A method for generating a neural network structure, comprising:
 - sampling the neural network structure to generate a network block, the network block comprising at least one network layer;
 - constructing a sampling neural network based on the network block;
 - training the sampling neural network based on sample data, and obtaining an accuracy corresponding to the sampling neural network; and
 - in response to that the accuracy does not meet a preset condition, regenerating a new network block according to the accuracy until a sampling neural network constructed by the new network block meets the preset condition, and using the sampling neural network meeting the preset condition as a target neural network.
2. The method according to claim 1, wherein the sampling the neural network structure to generate a network block comprises:
 - sampling the neural network structure based on a probability vector to obtain at least one type of network layers, network parameters corresponding to the at least one type of network layers, and a connection mode between the at least one type of network layers, and stacking the at least one type of network layers to obtain a network block.
3. The method according to claim 1 or 2, wherein the network layer comprises at least one of:
 - a convolution layer, a max pooling layer, an average pooling layer, an identity layer, an elemental add layer, or a contact layer.
4. The method according to claim 2 or 3, wherein the stacking the at least one type of network layers to obtain a network block comprises:
 - analyzing the network layers obtained by sampling to obtain at least one type of the network layers, a number of at least one type of the network layers, and a connection sequence of at least one type of the network layers; and
 - connecting the obtained network layers, according to the connection sequence corresponding to the at least one type of network layers through a connection

relationship between the at least one type of network layers to obtain a network block.

5. The method according to any one of claims 1-4, wherein the regenerating a new network block according to the accuracy comprises:

adjusting the probability vector of sampling the neural network structure according to the accuracy corresponding to the sampling neural network, and sampling the neural network structure through the adjusted probability vector, to generate a new network block.

6. The method according to any one of claims 1-5, wherein the training the sampling neural network based on sample data comprises:

processing the sample data by using the sampling neural network to obtain a processing result;

calculating an error value between supervision information marked in the sample data and the processing result by using a loss function, the sample data being marked with the supervision information; and

training the sampling neural network based on the error value until the sampling neural network meets a convergence condition.

7. The method according to claim 6, wherein the training the sampling neural network based on the error value until the sampling neural network meets a convergence condition comprises:

in response to that a convergence rate of the loss function is greater than or equal to a preset value, adjusting parameters in the sampling neural network through a reverse gradient algorithm according to the obtained error value to obtain an adjusted sampling neural network until the convergence rate of the loss function corresponding to the adjusted neural network is smaller than the preset value.

8. The method according to claim 6, wherein the training the sampling neural network based on the error value until the sampling neural network meets a convergence condition comprises:

in response to that a frequency of calculating the error value using the loss function is smaller than a preset value, adjusting parameters in the sampling neural

network through a reverse gradient algorithm according to the obtained error value to obtain an adjusted sampling neural network, and adding the frequency of calculating the error value using the loss function by 1;

until the frequency of calculating the error value using the loss function corresponding to the adjusted neural network is greater than or equal to the preset value.

9. The method according to any one of claims 6-8, wherein the obtaining an accuracy corresponding to the sampling neural network comprises:

calculating a predictive accuracy of the sampling neural network according to the error value of the trained sampling neural network; and

subtracting a preset proportion of network computing complexity and a preset proportion of network density based on the predictive accuracy to obtain an accuracy corresponding to the sampling neural network, the network computing complexity and the network density corresponding to the sampling neural network.

10. The method according to any one of claims 1-9, wherein the sampling the neural network structure to generate a network block comprises:

sampling the neural network structure to generate N network blocks, N being an integer greater than zero;

correspondingly, the constructing a sampling neural network based on the network block comprises:

respectively constructing N sampling neural networks based on the N network blocks;

correspondingly, the training the sampling neural network based on sample data, and obtaining an accuracy corresponding to the sampling neural network comprise:

respectively training the N sampling neural networks based on sample data, and obtaining N accuracies corresponding to the N sampling neural networks; and

correspondingly, in response to that the accuracy does not meet a preset condition, regenerating a new network block according to the accuracy until a sampling neural network constructed by the new network block meets the preset condition, and using the sampling neural network meeting the preset condition as a

target neural network comprise:

in response to that there is no accuracy among the N accuracies that meets a preset condition, respectively regenerating N new network blocks according to the N accuracies until there is a neural network among the N sampling neural networks respectively constructed by the N new network blocks that meets the preset condition, and using the sampling neural network meeting the preset condition as a target neural network.

11. The method according to claim 1, wherein after obtaining an accuracy corresponding to the sampling neural network, the method further comprises:

in response to that the accuracy meets the preset condition, using the sampling neural network meeting the preset condition as the target neural network;

or,

in response to that the accuracy does not meet the preset condition and the sampling for the neural network structure has reached a preset frequency, using a current sampling neural network as the target neural network.

12. An apparatus for generating a neural network structure, comprising:

a sampling unit, configured to sample the neural network structure to generate a network block, the network block comprising at least one network layer;

a network construction unit, configured to construct a sampling neural network based on the network block;

a training unit, configured to train the sampling neural network based on sample data, and obtaining an accuracy corresponding to the sampling neural network;
and

a network output unit, configured to, in response to that the accuracy does not meet a preset condition, regenerate a new network block according to the accuracy until a sampling neural network constructed by the new network block meets the preset condition, and use the sampling neural network meeting the preset condition as a target neural network.

13. The apparatus according to claim 12, wherein the sampling unit comprises:

a probability sampling module, configured to sample the neural network

structure based on a probability vector to obtain at least one type of network layers, network parameters corresponding to the at least one type of network layers, and a connection mode between the at least one type of network layers; and

a network block generation module, configured to stack the at least one type of network layers to obtain a network block.

14. The apparatus according to claim 12 or 13, wherein the network layer comprises at least one of:

a convolution layer, a max pooling layer, an average pooling layer, an identity layer, an elemental add layer, or a contact layer.

15. The apparatus according to claim 13 or 14, wherein the network block generation module is specifically configured to analyze the network layers obtained by sampling to obtain at least one type of the network layers, a number of at least one type of the network layers, and a connection sequence of at least one type of the network layers; and connect the obtained network layers according to the connection sequence corresponding to the at least one type of network layers through a connection relationship between the at least one type of network layers to obtain a network block.

16. The apparatus according to any one of claims 12-15, wherein the network output unit is specifically configured to adjust the probability vector of sampling the neural network structure according to the accuracy corresponding to the sampling neural network, and sample the neural network structure through the adjusted probability vector, to generate a new network block.

17. The apparatus according to any one of claims 12-16, wherein the training unit comprises:

a network processing module, configured to process the sample data by using the sampling neural network to obtain a processing result;

an error calculation module, configured to calculate an error value between supervision information marked in the sample data and the processing result by using a loss function, the sample data being marked with the supervision information; and

a network training module, configured to train the sampling neural network

based on the error value until the sampling neural network meets a convergence condition.

18. The apparatus according to claim 17, wherein the network training module is specifically configured to:

in response to that a convergence rate of the loss function is greater than or equal to a preset value, adjust parameters in the sampling neural network through a reverse gradient algorithm according to the obtained error value to obtain an adjusted sampling neural network until the convergence rate of the loss function corresponding to the adjusted neural network is smaller than the preset value.

19. An electronic device, comprising: a memory used for storing executable instructions; and

a processor, configured to communicate with the memory to execute the executable instructions to complete operations of the method for generating a neural network structure according to any one of claims 1-11.

20. A computer storage medium for storing computer readable instructions, wherein the instructions are executed to implement operations of the method for generating a neural network structure according to any one of claims 1-11.