CLAIMS

1. A model providing method including a neural network search method, wherein the method comprises:

receiving a performance requirement sent by a device side, wherein the performance requirement indicates a performance requirement of a neural network;

obtaining, from a plurality of candidate neural networks based on the performance requirement, a target neural network that meets the performance requirement, wherein at least one candidate neural network in the plurality of candidate neural networks comprises a target transformer layer, the target transformer layer comprises a target attention head, the target attention head comprises a plurality of operators, and the plurality of operators are obtained by sampling a plurality of candidate operators comprised in a first search space; and

sending the target neural network to the device side for data processing.

2. The method according to claim 1, wherein the target attention head is constructed based on the plurality of operators and an arrangement relationship between the plurality of operators, and the arrangement relationship between the plurality of operators is determined in a sampling manner.

3. The method according to claim 1 or 2, wherein the target attention head further comprises a first linear transformation layer, the first linear transformation layer is used to process an input vector of the target attention head by using a target transformation matrix, and the plurality of operators are used to perform an operation on a data processing result of the first linear transformation layer.

4. The method according to claim 3, wherein the target transformation matrix comprises only X transformation matrices, X is a positive integer less than or equal to 4, and a quantity of X is determined in a sampling manner.

5. The method according to any one of claims 1 to 4, wherein sizes of the input vector of the target attention head and an output vector of the target attention head are the same.

6. The method according to any one of claims 1 to 5, wherein a quantity of operators comprised in the target attention head is less than a preset value.

7. The method according to any one of claims 1 to 6, wherein the at least one candidate neural network comprises a plurality of network layers connected in series, the plurality of network layers comprise the target transformer layer, and a location of the target transformer layer in the plurality of network layers is determined in a sampling manner.

8. The method according to any one of claims 1 to 7, wherein the at least one candidate neural network comprises the plurality of network layers connected in series, the plurality of network layers comprise the target transformer layer and a target network layer, and the target network

layer comprises a convolutional layer.

9. The method according to any one of claims 1 to 8, wherein in a possible implementation, the performance requirement comprises at least one of the following: data processing precision, a model size, and an implemented task type.

10. The method according to any one of claims 1 to 9, wherein the target attention head is constructed based on the plurality of operators and an arrangement relationship between the plurality of operators, and the arrangement relationship between the plurality of operators is determined in a sampling manner.

11. A model providing apparatus, wherein the apparatus comprises:

a receiving module, configured to receive a performance requirement sent by a device side, wherein the performance requirement indicates a performance requirement of a neural network;

an obtaining module, configured to obtain, from a plurality of candidate neural networks based on the performance requirement, a target neural network that meets the performance requirement, wherein at least one candidate neural network in the plurality of candidate neural networks comprises a target transformer layer, the target transformer layer comprises a target attention head, the target attention head comprises a plurality of operators, and the plurality of operators are obtained by sampling a plurality of candidate operators comprised in a first search space; and

a sending module, configured to send the target neural network to the device side for data processing.

12. The apparatus according to claim 11, wherein the target attention head is constructed based on the plurality of operators and an arrangement relationship between the plurality of operators, and the arrangement relationship between the plurality of operators is determined in a sampling manner.

13. The apparatus according to claim 11 or 12, wherein the target attention head further comprises a first linear transformation layer, and the first linear transformation layer is used to process an input vector of the target attention head by using a target transformation matrix; and the plurality of operators are used to perform an operation on a data processing result of the first linear transformation layer.