CLAIMS

1. A computer-implemented method for positioning a description statement in an image, comprising:

performing analysis processing on a to-be-analyzed description statement and a to-be-analyzed image to obtain a plurality of statement attention weights of the to-be-analyzed description statement and a plurality of image attention weights of the to-be-analyzed image;

obtaining a plurality of first matching scores based on the plurality of statement attention weights and a subject feature, a location feature and a relationship feature of the to-be-analyzed image, wherein the to-be-analyzed image comprises a plurality of objects, a subject object is an object with a highest attention weight in the plurality of objects, the subject feature is a feature of the subject object, the location feature is a location feature of the plurality of objects, and the relationship feature is a relationship feature between the plurality of objects;

obtaining a second matching score between the to-be-analyzed description statement and the to-be-analyzed image based on the plurality of first matching scores and the plurality of image attention weights; and

determining a positioning result of the to-be-analyzed description statement in the to-be-analyzed image based on the second matching score,

wherein the method further comprises: obtaining the plurality of statement attention weights of the to-be-analyzed description statement and the plurality of image attention weights of the to-be-analyzed image by using a neural network,

wherein the method further comprises: before the obtaining the plurality of statement attention weights of the to-be-analyzed description statement and the plurality of image attention weights of the to-be-analyzed image by using a neural network, training the neural network by using a sample set, wherein the sample set comprises a plurality of positive sample pairs and a plurality of negative sample pairs,

wherein each positive sample pair comprises a first sample image and a first sample description statement of the first sample image, and

wherein each negative sample pair comprises the first sample image and a second sample description statement obtained after a word is removed from the first sample description statement, or comprises the first sample description statement and a second sample image obtained after a region with a highest image attention weight is removed from the first sample

image.

2. The computer-implemented method according to claim 1, wherein the performing analysis processing on a to-be-analyzed description statement and a to-be-analyzed image to obtain a plurality of statement attention weights of the to-be-analyzed description statement and a plurality of image attention weights of the to-be-analyzed image comprises:

performing feature extraction on the to-be-analyzed image to obtain an image feature vector of the to-be-analyzed image;

performing feature extraction on the to-be-analyzed description statement to obtain word embedding vectors of a plurality of words of the to-be-analyzed description statement; and

obtaining the plurality of statement attention weights of the to-be-analyzed description statement and the plurality of image attention weights of the to-be-analyzed image based on the image feature vector and the word embedding vectors of the plurality of words.

3. The computer-implemented method according to claim 1, wherein the plurality of statement attention weights comprises a statement subject weight, a statement location weight and a statement relationship weight;

the neural network comprises an image attention network;

the image attention network comprises a subject network, a location network and a relationship network;

the plurality of first matching scores comprises a subject matching score, a location matching score and a relationship matching score,

wherein the obtaining a plurality of first matching scores based on the plurality of statement attention weights and a subject feature, a location feature and a relationship feature of the to-be-analyzed image comprises:

inputting the statement subject weight and the subject feature into the subject network for processing to obtain the subject matching score;

inputting the statement location weight and the location feature into the location network for processing to obtain the location matching score; and

inputting the statement relationship weight and the relationship feature into the relationship network for processing to obtain the relationship matching score.

4. The computer-implemented method according to claim 3, wherein the plurality of image attention weights comprises a subject object weight, an object location weight and an object relationship weight,

wherein the obtaining a second matching score between the to-be-analyzed description statement and the to-be-analyzed image based on the plurality of first matching scores and the plurality of image attention weights comprises:

performing weighted averaging on the subject matching score, the location matching score and the relationship matching score based on the subject object weight, the object location weight and the object relationship weight to determine the second matching score.

5. The computer-implemented method according to claim 1, further comprising:

inputting the to-be-analyzed image into a feature extraction network for processing to obtain the subject feature, the location feature and the relationship feature.

6. The computer-implemented method according to claim 1, wherein the determining a positioning result of the to-be-analyzed description statement in the to-be-analyzed image based on the second matching score comprises:

responsive to the second matching score being greater than or equal to a preset threshold, determining an image region of the subject object as a positioning location of the to-be-analyzed description statement.

7. The computer-implemented method according to claim 1, wherein the neural network comprises a language attention network, wherein the method further comprises:

inputting the first sample description statement and the first sample image in the positive sample pair into the language attention network to obtain attention weights of a plurality of words of the first sample description statement;

replacing a word with a highest attention weight in the first sample description statement with a predetermined identifier to obtain the second sample description statement; and

using the first sample image and the second sample description statement as the negative sample pair.

8. The computer-implemented method according to claim 1, wherein the neural network comprises an image attention network, wherein the method comprises:

inputting the first sample description statement and the first sample image in the positive sample pair into the image attention network to obtain an attention weight of the first sample image;

removing an image region with a highest attention weight from the first sample image to obtain the second sample image; and

using the second sample image and the first sample description statement as the negative sample pair.

9. The computer-implemented method according to claim 1, wherein the training the neural network by using a sample set comprises:

determining an overall loss of the neural network based on a first loss and a second loss of the neural network; and

training the neural network based on the overall loss.

10. The computer-implemented method according to claim 9, further comprising: before the determining an overall loss of the neural network based on a first loss and a second loss of the neural network, obtaining the first loss,

wherein the operation of obtaining the first loss comprises:

inputting a first sample image and a first sample description statement in a same positive sample pair into the neural network for processing to obtain a first training score;

inputting a first sample image and a first sample description statement in different positive sample pairs into the neural network for processing to obtain a second training score; and

obtaining the first loss based on a plurality of first training scores and a plurality of second training scores.

11. The computer-implemented method according to claim 9, further comprising: before the determining an overall loss of the neural network based on a first loss and a second loss of the neural network, obtaining the second loss,

wherein the step of obtaining the second loss comprises:

inputting a second sample image and a first sample description statement in a same negative sample pair into the neural network for processing to obtain a third training score;

inputting a second sample image and a first sample description statement in different negative sample pairs into the neural network for processing to obtain a fourth training score;

inputting a first sample image and a second sample description statement in a same negative sample pair into the neural network for processing to obtain a fifth training score;

inputting a first sample image and a second sample description statement in different negative sample pairs into the neural network for processing to obtain a sixth training score; and

obtaining the second loss based on a plurality of third training scores, a plurality of fourth training scores, a plurality of fifth training scores and a plurality of sixth training scores.

12. The computer-implemented method according to claim 9, wherein the determining an overall loss of the neural network based on a first loss and a second loss of the neural network comprises:

performing weighted superposition on the first loss and the second loss to obtain the overall loss of the neural network.

13. An apparatus for positioning a description statement in an image, comprising:

a memory storing processor-executable instructions; and

a processor arranged to execute the stored processor-executable instructions to perform operations of:

performing analysis processing on a to-be-analyzed description statement and a

to-be-analyzed image to obtain a plurality of statement attention weights of the to-be-analyzed description statement and a plurality of image attention weights of the to-be-analyzed image;

obtaining a plurality of first matching scores based on the plurality of statement attention weights and a subject feature, a location feature and a relationship feature of the to-be-analyzed image, wherein the to-be-analyzed image comprises a plurality of objects, a subject object is an object with a highest attention weight in the plurality of objects, the subject feature is a feature of the subject object, the location feature is a location feature of the plurality of objects, and the relationship feature is a relationship feature between the plurality of objects;

obtaining a second matching score between the to-be-analyzed description statement and the to-be-analyzed image based on the plurality of first matching scores and the plurality of image attention weights; and

determining a positioning result of the to-be-analyzed description statement in the to-be-analyzed image based on the second matching score,

wherein the processor is arranged to execute the stored processor-executable instructions to further perform an operation of: obtaining the plurality of statement attention weights of the to-be-analyzed description statement and the plurality of image attention weights of the to-be-analyzed image by using a neural network,

wherein the processor is arranged to execute the stored processor-executable instructions to further perform an operation of: before the obtaining the plurality of statement attention weights of the to-be-analyzed description statement and the plurality of image attention weights of the to-be-analyzed image by using a neural network, training the neural network by using a sample set, wherein the sample set comprises a plurality of positive sample pairs and a plurality of negative sample pairs,

wherein each positive sample pair comprises a first sample image and a first sample description statement of the first sample image, and

wherein each negative sample pair comprises the first sample image and a second sample description statement obtained after a word is removed from the first sample description statement, or comprises the first sample description statement and a second sample image obtained after a region with a highest image attention weight is removed from the first sample image.

14. The apparatus according to claim 13, wherein the performing analysis processing on a to-be-analyzed description statement and a to-be-analyzed image to obtain a plurality of statement attention weights of the to-be-analyzed description statement and a plurality of image attention weights of the to-be-analyzed image comprises:

performing feature extraction on the to-be-analyzed image to obtain an image feature vector of the to-be-analyzed image;

performing feature extraction on the to-be-analyzed description statement to obtain word embedding vectors of a plurality of words of the to-be-analyzed description statement; and

obtaining the plurality of statement attention weights of the to-be-analyzed description statement and the plurality of image attention weights of the to-be-analyzed image based on the image feature vector and the word embedding vectors of the plurality of words.

15. The apparatus according to claim 13, wherein the plurality of statement attention weights comprises a statement subject weight, a statement location weight and a statement relationship weight, the neural network comprises an image attention network, the image attention network comprises a subject network, a location network and a relationship network, and the plurality of first matching scores comprises a subject matching score, a location matching score and a relationship matching score,

wherein the obtaining a plurality of first matching scores based on the plurality of statement attention weights and a subject feature, a location feature and a relationship feature of the to-be-analyzed image comprises:

inputting the statement subject weight and the subject feature into the subject network for processing to obtain the subject matching score;

inputting the statement location weight and the location feature into the location network for processing to obtain the location matching score; and

inputting the statement relationship weight and the relationship feature into the relationship network for processing to obtain the relationship matching score.16. The apparatus according to claim 15, wherein the plurality of image attention weights comprises a subject object weight, an object location weight and an object relationship weight,

wherein the obtaining a second matching score between the to-be-analyzed description statement and the to-be-analyzed image based on the plurality of first matching scores and the

plurality of image attention weights comprises:

performing weighted averaging on the subject matching score, the location matching score and the relationship matching score based on the subject object weight, the object location weight and the object relationship weight to determine the second matching score.

17. A non-transitory computer readable storage medium having stored thereon computer program instructions that, when executed by a processor, cause the processor to perform operations of a computer-implemented method for positioning a description statement in an image, the method comprising:

performing analysis processing on a to-be-analyzed description statement and a to-be-analyzed image to obtain a plurality of statement attention weights of the to-be-analyzed description statement and a plurality of image attention weights of the to-be-analyzed image;

obtaining a plurality of first matching scores based on the plurality of statement attention weights and a subject feature, a location feature and a relationship feature of the to-be-analyzed image, wherein the to-be-analyzed image comprises a plurality of objects, a subject object is an object with a highest attention weight in the plurality of objects, the subject feature is a feature of the subject object, the location feature is a location feature of the plurality of objects, and the relationship feature is a relationship feature between the plurality of objects;

obtaining a second matching score between the to-be-analyzed description statement and the to-be-analyzed image based on the plurality of first matching scores and the plurality of image attention weights; and

determining a positioning result of the to-be-analyzed description statement in the to-be-analyzed image based on the second matching score,

wherein the method further comprises: obtaining the plurality of statement attention weights of the to-be-analyzed description statement and the plurality of image attention weights of the to-be-analyzed image by using a neural network,

wherein the method further comprises: before the obtaining the plurality of statement attention weights of the to-be-analyzed description statement and the plurality of image attention weights of the to-be-analyzed image by using a neural network, training the neural network by

using a sample set, wherein the sample set comprises a plurality of positive sample pairs and a plurality of negative sample pairs,

wherein each positive sample pair comprises a first sample image and a first sample description statement of the first sample image, and

wherein each negative sample pair comprises the first sample image and a second sample description statement obtained after a word is removed from the first sample description statement, or comprises the first sample description statement and a second sample image obtained after a region with a highest image attention weight is removed from the first sample image.

18. The non-transitory computer readable storage medium according to claim 17, wherein the performing analysis processing on a to-be-analyzed description statement and a to-be-analyzed image to obtain a plurality of statement attention weights of the to-be-analyzed description statement and a plurality of image attention weights of the to-be-analyzed image comprises:

performing feature extraction on the to-be-analyzed image to obtain an image feature vector of the to-be-analyzed image;

performing feature extraction on the to-be-analyzed description statement to obtain word embedding vectors of a plurality of words of the to-be-analyzed description statement; and

obtaining the plurality of statement attention weights of the to-be-analyzed description statement and the plurality of image attention weights of the to-be-analyzed image based on the image feature vector and the word embedding vectors of the plurality of words.

19. The non-transitory computer readable storage medium according to claim 17, wherein the plurality of statement attention weights comprises a statement subject weight, a statement location weight and a statement relationship weight;

the neural network comprises an image attention network;

the image attention network comprises a subject network, a location network and a relationship network;

the plurality of first matching scores comprises a subject matching score, a location matching score and a relationship matching score,

wherein the obtaining a plurality of first matching scores based on the plurality of

statement attention weights and a subject feature, a location feature and a relationship feature of the to-be-analyzed image comprises:

inputting the statement subject weight and the subject feature into the subject network for processing to obtain the subject matching score;

inputting the statement location weight and the location feature into the location network for processing to obtain the location matching score; and

inputting the statement relationship weight and the relationship feature into the relationship network for processing to obtain the relationship matching score.

20. The non-transitory computer readable storage medium according to claim 19, wherein the plurality of image attention weights comprises a subject object weight, an object location weight and an object relationship weight,

wherein the obtaining a second matching score between the to-be-analyzed description statement and the to-be-analyzed image based on the plurality of first matching scores and the plurality of image attention weights comprises:

performing weighted averaging on the subject matching score, the location matching score and the relationship matching score based on the subject object weight, the object location weight and the object relationship weight to determine the second matching score.