

NEURAL NETWORK MODEL ADJUSTMENT SYSTEM AND METHOD, AND DEVICE

5 The present application claims the priority to a Chinese patent application No. 202011431789.8 filed with the China National Intellectual Property Administration on December 7, 2020 and entitled “Neural Network Model Adjustment system, Method, and Device”, which is incorporated herein by reference in its entirety.

10 **Technical Field**

The present application relates to the technical field of machine learning, and in particular to a neural network model adjustment system, method, and device.

Background

15 Machine learning has been widely applied in various fields, for example, image classification, target detection, natural language processing, etc. During machine learning, sample images can be used to train a neural network. The process of training is a process of adjusting iteratively network parameters of the neural network. A neural network model obtained after training is completed can be configured for image classification, target detection, natural
20 language processing, etc.

In relevant machine learning solutions, generally a neural network is trained using sample images collected for a same scenario. As a result, a neural network model obtained after training is completed has relatively poor generalizability. Generalizability can be understood as the capacity of a neural network model to adapt to different scenarios. A neural network model with
25 relatively poor generalizability has a relatively low output accuracy in scenarios to which it is not adapted.

Using a target detection model as an example, a target detection model can recognize various targets in an image such as a person, a vehicle, a traffic sign, etc. For example, a neural

network is trained using sample images collected in a traffic scenario, to obtain a target detection model. When target detection is performed on images collected in the traffic scenario using the target detection model, the accuracy is relatively high; but when target detection is performed on images collected in a building scenario using the target detection model, the accuracy is relatively low.

Summary

An objective of embodiments of the present application is to provide a neural network model adjustment system, method, and device, so as to increase a model's output accuracy.

To achieve the above-described objective, an embodiment of the present application provides a neural network model adjustment system, including a server and a plurality of clients, wherein each of the clients corresponds respectively to a different scenario; wherein,

the server is configured for sending a generalized model to each of the clients;

the clients are configured for obtaining sample images collected in scenarios respectively corresponding to the clients as private sample images; adjusting the generalized model using the private sample images to obtain a data processing model; and sending the private sample images to the server;

the server is further configured for adjusting the generalized model based on the private sample images sent by each of the clients, and sending the adjusted generalized model to each of the clients.

Optionally, the server is further configured for sending an encryption model to each of the clients;

the clients are specifically configured for encrypting their own private sample images using the encryption model to obtain encrypted data; and sending the encrypted data to the server;

the server is specifically configured for adjusting the generalized model based on the encrypted data sent by each of the clients.

Optionally, the encryption model includes a feature extraction model;

the clients are specifically configured for performing feature extraction on the private sample images using the feature extraction model to obtain encrypted data.

Optionally, the encryption model further includes a compression encoding model;

5 the clients are specifically configured for performing feature extraction on the private sample images using the feature extraction model to obtain features of the private sample images; performing compression encoding on the features of the private sample images using the compression encoding model to obtain encoded features; sending the encoded features to the server as encrypted data;

10 the server is specifically configured for decoding the encoded features using a decoding model to obtain decoded features, and adjusting the generalized model based on the decoded features.

Optionally, the generalized model includes a low-level model and a high-level model, wherein the low-level model is a partial model close to an input layer, and the high-level model is a partial model close to an output layer; the feature extraction model is the low-level model.

15 To achieve the above-described objective, an embodiment of the present application further provides neural network model adjustment method, which is applied to a first client, wherein the method includes:

20 receiving a generalized model sent by a server, wherein the generalized model is obtained by the server adjusting the generalized model based on encrypted data sent by a plurality of clients, and each of the clients corresponds respectively to a different scenario;

obtaining sample images collected in a scenario corresponding to the first client as private sample images;

adjusting the generalized model using the private sample images to obtain a data processing model;

25 sending the private sample images to the server.

Optionally, the method further includes:

receiving an encryption model sent by the server;

sending the private sample images to the server includes:

encrypting the private sample images using the encryption model to obtain encrypted data;
and sending the encrypted data to the server.

5 Optionally, the encryption model includes a feature extraction model; encrypting the private sample images using the encryption model to obtain the encrypted data, includes:

performing feature extraction on the private sample images using the feature extraction model to obtain the encrypted data.

10 Optionally, the encryption model further includes a compression encoding model; performing feature extraction on the private sample images using the feature extraction model to obtain the encrypted data, includes:

performing feature extraction on the private sample images using the feature extraction model to obtain features of the private sample images;

performing compression encoding on the features of the private sample images using the compression encoding model to obtain encoded features as the encrypted data;

15 Optionally, the method further includes:

after receiving the generalized model, determining whether the generalized model satisfies a preset performance condition;

if the generalized model satisfies the preset performance condition, determining the generalized model as the data processing model;

20 otherwise, executing the step of obtaining the sample images collected in the scenario corresponding to the first client as private sample images.

To achieve the above-described objective, an embodiment of the present application further provides a neural network model adjustment method, applied to a server having stored therein a generalized model, wherein the method includes:

25 receiving private sample images sent by each of clients, wherein each of the clients corresponds respectively to a different scenario, and the private sample images are sample images collected in scenarios respectively corresponding to each of the clients;

adjusting the generalized model stored in the server based on the private sample images sent by each of the clients;

sending the adjusted generalized model to each of the clients, so that each of the clients adjusts the generalized model using its own private sample images.

5 Optionally, an encryption model is further stored in the server; the method further includes: sending the encryption model to each of the clients; receiving private sample images sent by each of the clients includes: receiving encrypted data sent by each of the clients, wherein the encrypted data is data obtained by the clients encrypting the sample images collected in the clients' own corresponding scenarios using the encryption model.

10 Optionally, the generalized model includes a low-level model and a high-level model, wherein, the low-level model is a partial model close to an input layer, and the high-level model is a partial model close to an output layer;

the encryption model includes a feature extraction model, and the feature extraction model is the low-level model; the encrypted data are data obtained by the clients performing feature
15 extraction on sample images collected in the clients' own corresponding scenarios using the feature extraction model;

adjusting the generalized model stored in the server based on the private sample images sent by each of the clients includes: inputting the encrypted data into the high-level model, and adjusting the high-level model based on an output result of the high-level model.

20 Optionally, the generalized model includes a low-level model and a high-level model, wherein the low-level model is a partial model close to an input layer, and the high-level model is a partial model close to an output layer;

the encryption model includes a feature extraction model and a compression encoding model, and the feature extraction model is the low-level model;

25 the encrypted data are data obtained by the clients performing, sequentially, feature extraction and compression encoding on sample images collected in the clients' own corresponding scenarios using the feature extraction model and the compression encoding model;

adjusting the generalized model stored in the server based on the private sample images sent by each of the clients includes: decoding the encrypted data by means of the decoding model to obtain decoded features; inputting the decoded features into the high-level model, and adjusting the high-level model based on an output result of the high-level model.

5 To achieve the above-described objective, an embodiment of the present application further provides an electronic device, wherein it includes a processor and a memory;

the memory is configured for storing a computer program;

the processor is configured for, when executing the program stored in the memory, carrying out any neural network model adjustment method.

10 To achieve the above-described objective, an embodiment of the present application further provides a computer readable storage medium having stored therein a computer program, wherein the computer program, when executed by a processor, causes the processor to carry out any neural network model adjustment method.

15 To achieve the above-described objective, an embodiment of the present application further provides a computer program product containing instructions which, when running on a computer, causes the computer to carry out any neural network model adjustment method.

By applying embodiments shown in the present application, a server sends a generalized model to each of clients corresponding respectively to a different scenario; the clients obtain sample images collected in corresponding scenarios as private sample images; the generalized
20 model is adjusted using the private sample images to obtain a data processing model; the private sample images are sent to the server, and the server adjusts the generalized model based on the private sample images.

It can be seen that in the present solutions, in a first aspect, the server adjusts the generalized model based on sample images sent by each of the clients corresponding respectively
25 to a different scenario, in other words, the server adjusts the generalized model based on multi-scenario images, and therefore, the generalized model can adapt to multiple scenarios, which increases the output accuracy of the generalized model in different scenarios; in a second aspect, the clients adjust the generalized model using the private sample images for their own scenarios to obtain a data processing model, wherein the data processing model can have at the

same time generalizability that adapts to multiple scenarios and stronger adaptability to specific scenarios, in other words, the data processing model has at the same time generalizability and specificity, which increases the output accuracy of the data processing model in different scenarios. Of course, an implementation of any product or method of the present application
5 does not need to achieve all the above-described advantages at the same time.

Brief Description of the Drawings

In order to describe the technical solutions of embodiments of the present application or of the prior art more clearly, a simple introduction of the drawings required in the description of the
10 embodiments and of the prior art will be given. Obviously, the drawings described below are just for some embodiments of the present application and other drawings can be obtained by those of ordinary skills in the art based on these drawings without any creative effort.

Fig. 1 is a first schematic structural view of a neural network model adjustment system provided by an embodiment of the present application;

15 Fig. 2a is a first schematic view of interactions between a server and a client provided by an embodiment of the present application;

Fig. 2b is a second schematic view of interactions between a server and a client provided by an embodiment of the present application;

20 Fig. 3 is a second schematic structural view of a neural network model adjustment system provided by an embodiment of the present application;

Fig. 4 is a third schematic view of interactions between a server and a client provided by an embodiment of the present application;

Fig. 5 is a schematic flow chart of a neural network model adjustment method applied to a first client provided by an embodiment of the present invention;

25 Fig. 6 is a schematic flow chart of a neural network model adjustment method applied to a server provided by an embodiment of the present invention;

Fig. 7 is a schematic structural view of an electronic device provided by an embodiment of the present invention.

Detailed Description

In order to make the objects, technical solutions and advantages more clearly, the present application will be further described in detail below with reference to the appended drawings and
5 embodiments. Apparently, the described embodiments are only some of the embodiments of the present disclosure, not all of them. All other embodiments obtained by those of ordinary skill in the art based on the embodiments in the present application without creative efforts fall within the protection scope of the present application.

To achieve the above-described objective, embodiments of the present application provide a
10 neural network model adjustment system, method, device, a computer readable storage medium, and a computer program product. The neural network model adjustment system is described in detail first below.

Fig. 1 is a first schematic structural view of a neural network model adjustment system provided by an embodiment of the present application, including: a server 100 and a plurality of
15 clients (client 1, . . . , client N, N representing a positive integer larger than 1) 200, wherein the server 100 are communicatively connected with each of the clients 200, wherein each of the clients 200 corresponds respectively to a different scenario.

As an example, the server and clients in the embodiment of the present application can be hardware devices, for example, a server and user devices respectively, or software programs, for
20 example, server software and client software respectively. The specific forms of the server and clients are not limited.

Wherein, one can refer to Fig. 2a regarding interactions between the server and any client
200:

the server 100 sends a generalized model to each of the clients 200;

25 the clients 200 obtain sample images collected in scenarios respectively corresponding to the clients as private sample images, adjust the generalized model using the private sample images to obtain a data processing model, and send the private sample images to the server 100;

the server 100 adjusts the generalized model based on the private sample images sent by each of the clients 200, and sends the adjusted generalized model to each of the clients 200.

The various models mentioned in embodiments of the present application, such as the generalized model here, and the encryption model, the data processing model, the feature extraction model, the compression encoding model, the decoding model, etc. in the following contents, can all be understood as a neural network model, or a partial layer or level in a neural network model. The specific type of the neural network model is not limited, such as a deep neural network model, a convolutional neural network model, etc.

The neural network model can be an image classification model, a target detection model, a text recognition model, or a video processing model, etc. The specific uses of the neural network model are not specifically listed. Correspondingly, if the neural network model is an image classification model, then sample images are single-frame images or video images. If the neural network model is a text recognition model, then sample images are text images. If the neural network model is a video processing model, then sample images are video data. The type of sample images matches the type of the neural network model. They are not specifically listed here.

Each of the clients corresponds respectively to a different scenario. Private sample images of a client are sample images collected in a scenario corresponding to that client, which are sample images specific to that client.

The server 100 can continuously adjust the generalized model based on the private sample images sent by each of the clients. As each of the clients corresponds respectively to a different scenario, the generalized model is adjusted based on multi-scenario data. The generalized model can adapt to multiple scenarios and has relatively good generalizability.

The server 100 sends the generalized model to each of the clients 200. Each of the clients 200 makes adjustments based on the generalized model sent by the server 100 to obtain a data processing model. In practice, the data processing model can be configured for image classification, target detection, text recognition, video processing, etc. Its specific functions are not limited. The data processing model is obtained by adjustment based on the generalized model. Therefore, the data processing model also has relatively good generalizability.

In addition, as to each client, the client obtains sample images collected in its corresponding scenario, and uses the sample images to adjust the generalized model to obtain a data processing

model. In other words, the client adjusts the generalized model using private data of its own scenario. The data processing model obtained after adjustment has better capacity of processing private data of the client's corresponding scenario. It can be seen that the data processing model has both generalizability of adapting to various scenarios and stronger adaptability to a certain specific scenario. In other words, the data processing model has both versatility and specificity.

In one implementation, after receiving the generalized model sent by the server 100, the clients 200 can first determine whether the generalized model satisfies a preset performance condition; if it does, directly determine the generalized model as the data processing model; otherwise, execute the step of obtaining private sample images, and adjusting the generalized model using the private sample images to obtain the data processing model.

The preset performance condition can be set based on actual needs. Different clients' corresponding preset performance conditions can be different or identical. As an example, if the generalized model is a target detection model, then the preset performance condition can include any one or more of the following: undetected rate being lower than a first preset threshold, positive rate being higher than a second preset threshold, accuracy being higher than a third preset threshold, etc., which are not specifically listed. As another example, if the generalized model is a text recognition model, then the preset performance condition can be accuracy being higher than a fourth preset threshold, etc. Specific preset performance conditions are not limited.

In this implementation, if the generalized model satisfies a preset performance condition, then the generalized model can be directly applied. The generalized model is directly determined as the data processing model. In other words, the generalized model is directly applied for actual data processing. If the generalized model does not satisfy the preset performance condition, then the method proceeds to subsequent adjustment and other steps.

In one implementation, the server 100 can further send an encryption model to each of the clients 200; the clients 200 encrypt their own private sample images using the encryption model to obtain encrypted data, and send the encrypted data to the server 100; the server 100 adjusts the generalized model based on the encrypted data sent by each of the clients 200.

In this implementation, by encrypting the private sample images of the clients before sending them to the server, data safety is increased.

In this implementation, one can refer to Fig. 2b regarding interactions between the server and any client 200:

the server 100 sends the generalized model and the encryption model to each of the clients 200;

5 the clients 200 obtain sample images collected in scenarios respectively corresponding to the clients as private sample images; adjust the generalized model using the private sample images to obtain a data processing model; encrypt the private sample images using the encryption model to obtain encrypted data; and send the encrypted data to the server 100;

10 the server 100 adjusts the generalized model based on the encrypted data, and sends the adjusted generalized model to each of the clients 200.

As an example, the encryption model can be fixed. The server 100 can send the encryption model to each of the clients 200 only once. But the generalized model is not fixed. The server 100 continuously adjusts the generalized model based on the private sample images sent by each of the clients, and continuously sends the adjusted generalized model to each of the clients 200.
15 Wherein, the encryption model is configured for processing private sample images to obtain corresponding data, so as to achieve an objective of encrypting the private sample images.

In one implementation, the encryption model includes a feature extraction model. The clients 200 can perform feature extraction on the private sample images using the feature extraction model to obtain encrypted data.

20 In an example, the clients 200 perform feature extraction on the private sample images using the feature extraction model to obtain features of the private sample images and send the features as encrypted data to the server 100. In other words, in this example, feature extraction is used as a means of encryption. Feature extraction can be understood as lossy processing of sample images, thus achieving an effect of encrypting the sample images and ensuring the safety
25 of sample images. Even if an unauthorized user obtains features of the sample images, he/she cannot restore the sample images.

In this example, after receiving the features of the private sample images, the server 100 can directly use the features of the private sample images to adjust the generalized model.

In another example, the encryption model includes a feature extraction model and a compression encoding model.

The clients 200 can perform feature extraction on the private sample images using the feature extraction model to obtain features of the private sample images; perform compression
5 encoding on the features of the private sample images using the compression encoding model to obtain encoded features; and send the encoded features to the server as encrypted data;

the server 100 is further configured for decoding the encoded features using a decoding model to obtain decoded features, and adjusting the generalized model based on the decoded features.

10 In this embodiment, using feature extraction and feature encoding as a means of encryption is equivalent to perform two operations of lossy processing on the sample images. On one hand, it achieves an effect of encrypting the sample images, ensuring the safety of the sample images and making it impossible to restore the sample images even if an unauthorized user obtains encoded features. On the other hand, feature encoding can reduce data quantity, reduce
15 bandwidth consumption during data transfer, and also reduce memory consumption during subsequent data processing.

Alternatively, in another implementation, the clients can use another encryption algorithm model to encrypt sample images, and the server uses a corresponding decryption algorithm model to decrypt the encrypted data. The specific encryption and decryption algorithms are not
20 limited.

In the above-described embodiment in which the clients send encrypted data corresponding to private sample images to the server, as the encrypted data corresponding to private sample images can reflect features of the private sample images, an effect of sending the private sample images to the server can be achieved. As the private sample images are not directly sent to the
25 server, data safety is ensured at the same time.

As an example, in one implementation, the generalized model includes a low-level model and a high-level model. The low-level model is a partial model close to an input layer, and the high-level model is a partial model close to an output layer. The feature extraction model is the low-level model. Wherein, the classification of the low-level model and the high-level model can

be determined based on the actual structure of the generalized model.

For example, the low-level model can include a small number of convolutional layers, BN (Batch Normalization) layers, and pooling layers. The low-level layer is configured for performing feature extraction on input data.

5 In an above-described example, the server receives features of sample images sent by clients. The features are features output by the feature extraction model. In this implementation, the feature extraction model is a low-level model in the generalized model. Therefore, Features of sample images received by the server can be understood as features obtained by processing sample images by the low-level model of the generalized model. Thus, when the server adjusts
10 the generalized model based on encrypted data, features of the sample images can be directly input into a high-level model of the generalized model. It can be seen that by applying this implementation, the process of the server's adjustment on the generalized model can be simplified.

 In another above-described example, the server receives encoded features sent by clients.
15 The process of processing sample images by the clients can be expressed as: sample images → features output by the feature extraction model → encoded features output by the compression encoding model. The server decodes the encoded features. Decoding can be understood as a reverse process of encoding. The steps of encoding and decoding cancel each other. Thus, the encoded features are equivalent to features output by the feature extraction model. In this
20 implementation, the feature extraction model is a low-level model in the generalized model. Therefore, decoded features received by the server can be understood as features obtained by processing sample images by the low-level model of the generalized model. Thus, when the server adjusts the generalized model based on decoded features, features of the sample images can be directly input into a high-level model of the generalized model. It can be seen that by
25 applying this implementation, the process of the server's adjustment on the generalized model can be simplified.

 In this implementation, the feature extraction model is a low-level model of the generalized model. That is, the feature extraction model is part of the generalized model. Alternatively, in another case, the feature extraction model and the generalized model can be two independent
30 models, but the feature extraction model is consistent with the low-level model of the

generalized model in terms of the layering structure and the various parameters of each layer.

By way of example, there may be various situations for the adjustment of the generalized model by the server 100:

For example, there may be various situations for the timing of the adjustment of the generalized model by the server 100: in one situation, when data sent by the clients 200 and newly received by the server 100 reach a preset data quantity, the generalized model is adjusted, that is, when data sent by the clients 200 and received by the server 100 after adjusting the generalized model last time reach a preset data quantity, then the generalized model is adjusted once again; in another situation, the server 100 can adjust the generalized model once every preset time interval; in yet another situation, the server 100 can adjust the generalized model each time after receiving data sent by the clients 200. The specific timing of adjustment is not limited.

As another example, there may be various situations for data based on which the server 100 adjusts the generalized model:

in one situation, the server 100 can store data sent by each of the clients 200, and each time it adjusts the generalized model, it adjusts the generalized model based on all the stored data. Alternatively, in another situation, the generalized model is adjusted each time based on different data. For example, in the above-described example, the server 100 can adjust the generalized model based on newly received data sent by the clients 200. That is, the server 100 adjusts the generalized model based on data sent by the clients 200 and received after adjusting the generalized model last time. The data based on which adjustment is made is not limited.

The server 100 can accumulate more and more data sent by the clients 200 and continuously adjust the generalized model based on those data, and the performance of the generalized model becomes better and better. It can be seen that the neural network model adjustment system provided by embodiments of the present application has a self-learning function. The server 100 can send an adjusted generalized model to the clients 200 each time after adjusting the generalized model, and the clients 200 can make adjustments based on this generalized model to obtain a data processing model with a better performance.

A specific implementation is described below in reference to Figs. 3 to 4.

In reference to Fig. 3, the server 100 includes a generalized model, a feature extraction model, a compression encoding model, and a decoding model:

Generalized Model

5 The generalized model can be divided in two parts, i.e., a low-level model, and a high-level model. The low-level model is a part close to an input layer, and the high-level model is a part close to an output layer. The low-level model generally includes a small number of convolutional layers, BN layers, and pooling layers. A function of the low-level model is to perform feature extraction on input data.

10 The generalized model is shared in the entire system. The generalized model is capable of adapting to clients' corresponding scenarios and has relatively good generalizability. The generalized model is obtained by training encrypted data sent by clients 200. For example, in Fig. 3, each of the clients 200 sends encoded features to the server 100. The server 100 decodes the encoded features by means of the decoding model to obtain decoded features, and trains the
15 generalized model based on the decoded features. Subsequently, the server 100 can update the trained generalized model based on the encrypted data newly sent by the clients.

Feature Extraction Model

 The low-level model in the generalized model can be used as the feature extraction model. In one situation, data output by the feature extraction model can be used as encrypted data for
20 transfer. A feature extraction process is a lossy processing process. Therefore, by performing feature extraction on sample images before transfer, privacy of the sample images can be protected. Further, when training/updating the generalized model, data obtained after feature extraction can be directly sent into the high-level model of the generalized model.

 The inventor has discovered through studies that in a pre-trained neural network model,
25 fixing parameters of a low-level (or, lower level) model has very little effect on the overall performance of the neural network model. Therefore, in the entire system, parameters of the low-level model can be fixed. During the process of training/updating the generalized model by the server 100, and the process of adjusting the generalized model based on the private sample images by the clients 200, the parameters of the low-level model may not be adjusted, and only

parameters of the high-level model are adjusted.

Compression Encoding Model

A function of the compression encoding model is to perform compression encoding on features extracted by the low-level model. The process of compression encoding is a further
5 lossy processing process. The compression ratio of the compression encoding model can reach several hundreds or even several thousands. After being compressed at such a high ratio, the original sample images suffer further loss, and it is more difficult to restore the original sample images from encoded features. By performing feature extraction and compression encoding on sample images before transfer, privacy of the sample images can be further protected. In
10 addition, when training/updating the generalized model, decoded features can be directly sent into the high-level model of the generalized model.

Generally, data quantity after feature extraction of sample images will be several times or even several ten times larger than the data quantity of the original sample images. Directly transferring features extracted by the feature extraction model will take up a large amount of
15 bandwidth, and it will take the server a large amount of memory space to store features transferred by each of the clients by the server. By using the compression encoding model to perform compression encoding on the features before transfer, which not only saves bandwidth, but also saves memory space.

The compression encoding model can be trained by the server, and may not be adjusted any
20 more after completion of training.

Decoding Model

Due to the special nature of a neural network, the compression ratio of the compression encoding model can reach several hundreds or even several thousands. With such a compression ratio, the decoding model can reconstruct features that approximate those before the compression
25 encoding. That is, decoded features output by the decoding model approximate features output by the feature extraction model. The decoded features can be directly sent into the high-level model of the generalized model. The high-level model of the generalized model is adjusted based on the decoded features.

In reference to Fig. 3, the client 200 includes a generalized model, a feature extraction

model, a compression encoding model, and a private model. For ease of presentation, only one client is shown in Fig. 3. Other clients are similar.

The generalized model, feature extraction model, and compression encoding model in the client 200 are all sent by the server. The private model in the client 200 is obtained by the client 200 by adjusting the generalized model based on private data of the client 200. Each of the clients corresponds respectively to a different scenario. A client's own private data can be understood as sample images collected in the client's corresponding scenario.

To the client 200, the generalized model sent by the server 100 may or may not have sufficient generalizability in the client's corresponding scenario. If it does, then the client 200 can directly use this generalized model, or in other words, directly use the generalized model as its own private model. Otherwise, the client 200 can obtain its own private data and adjust the generalized model based on the private data, thus obtaining a private model adapted to its own scenario. In addition, encoded features are obtained after feature extraction and compression encoding on private data obtained by the client 200. The client 200 can send the encoded features to the server 100. The server 100 can decode the encoded features sent by each of the clients 200, and perform training based on the decoded features to obtain a generalized model with better versatility.

Wherein, each client 200's private data and private model are inaccessible to other clients or server.

In reference to Fig. 4, a flow process of execution in the system is described:

a server 100 sends a generalized model, a low-level model, and a compression encoding model to a client 200;

the client 200 evaluates whether the performance of the generalized model satisfies a preset performance condition, and if it does, then the procedure ends, otherwise, subsequent steps are performed;

the client 200 obtains its scenario's private data and adjusts the generalized model based on the private data, so as to obtain a private model the performance of which satisfies the preset performance condition;

the client 200 sends the private data into the feature extraction model for feature extraction, and then into the compression encoding model for compression encoding, and then send encoded features to the server 200;

5 the server 200 receives and stores encoded features sent by the client 200. When the scale of newly added encoded features reaches a certain scale, the server 200 performs joint training based on all the stored encoded features to obtain a new generalized model with better generalizability. The server 200 sends the new generalized model to all the clients 200.

To a client having private data and a private model, it can choose to make readjustment based on the new generalized model to obtain a private model with a better performance.

10 It can be seen that this implementation provides a neural network model adjustment system that is capable of collaborative learning. In a first aspect, the system has the capability of self learning. New clients can be added continuously into the system. With the continuous addition of new clients, the server will continuously accumulate more and more encoded features, and thus obtains by training a generalized model with better generalizability. In a second aspect, with the
15 continuous strengthening of the generalized model's generalizability, the probability that newly added clients encounter generalizability problems will continually decrease. Thus, human and material resources needed by various procedures such as data collection, data marking, and model training can be saved. In a third aspect, data transferred between a client and a server in the system are encoded features. Thus, the system's data safety can be provided. In a fourth
20 aspect, the system can use private data of different clients to together train a generalized model that has strong versatility and good generalizability. Compared to a solution of isolated data islands in which each client obtains private data by itself, calibrates data by itself, and trains a model by itself, the system does not need to perform this series of time-consuming and labor-intensive, repetitive labor.

25 Correspondingly to the above-described system embodiments, an embodiment of the present application further provides a neural network model adjustment method applied to a first client. The first client can be any client in a neural network model adjustment system. To distinguish an executing entity from other clients, the executing entity is referred to as the first client.

Fig. 5 is a schematic flow chart of a neural network model adjustment method applied to a first client provided by an embodiment of the present invention, including:

S501: receiving a generalized model sent by a server.

5 The generalized model is obtained by the server adjusting the generalized model based on encrypted data sent by a plurality of clients, and each of the clients corresponds respectively to a different scenario.

10 The various models mentioned in the embodiment of the present application, such as the generalized model here, and the encryption model, data processing model, feature extraction model, compression encoding model, decoding model, etc. in the following contents, can all be understood as a neural network model, or a partial layer or level in a neural network model. The specific type of the neural network model is not limited, such as a deep neural network model, a convolutional neural network model, etc.

15 The neural network model can be an image classification model, a target detection model, a text recognition model, or a video processing model, etc. The specific uses of the neural network model are not specifically listed. Correspondingly, if the neural network model is an image classification model, then sample images are single-frame images or video images. If the neural network model is a target detection model, then sample images are single-frame images or video images. If the neural network model is a text recognition model, then sample images are text images. If the neural network model is a video processing model, then sample images are video data. The type of sample images matches the type of the neural network model, and is not specifically listed here.

25 The server can continuously adjust the generalized model based on the private sample images sent by each of the clients. As each of the clients corresponds respectively to a different scenario, the generalized model is adjusted based on multi-scenario data. The generalized model can adapt to multiple scenarios and has relatively good generalizability. A client's private sample images are sample images collected in the client's corresponding scenario, which are sample images specific to the client.

In one implementation, after stepS501, it can be first determined whether the generalized model satisfies a preset performance condition; if it does, then the generalized model is

determined as the data processing model; otherwise, then step S502 is executed again.

The preset performance condition can be set based on actual needs. Different clients' corresponding preset performance conditions can be different or identical. As an example, if the generalized model is a target detection model, then the preset performance condition can include
5 any one or more of the following: undetected rate being lower than a first preset threshold, positive rate being higher than a second preset threshold, accuracy being higher than a third preset threshold, etc., which are not specifically listed. As another example, if the generalized model is a text recognition model, then the preset performance condition can be accuracy being higher than a fourth preset threshold, etc. Specific preset performance conditions are not limited.

10 In this implementation, if the generalized model satisfies a preset performance condition, then the generalized model can be directly applied. The generalized model is directly determined as the data processing model. In other words, the generalized model is directly applied for actual data processing. If the generalized model does not satisfy the preset performance condition, then the method proceeds to subsequent adjustment and other steps.

15 Alternatively, in another implementation, after step S501, step S502 can be directly executed.

S502: obtaining sample images collected in a scenario corresponding to the first client as private sample images.

20 In this embodiment, each of the clients corresponds respectively to a different scenario, and sample images collected by the first client in its corresponding scenario can be understood as private data specific to the first client.

S503: adjusting the generalized model using the private sample images to obtain a data processing model.

25 The first client makes adjustments based on the generalized model sent by the server to obtain a data processing model. In practice, the data processing model can be configured for image classification, target detection, text recognition, video processing, etc. Its specific functions are not limited. The data processing model is obtained by adjustment based on the generalized model. Therefore, the data processing model also has relatively good generalizability.

In addition, the first client obtains sample images collected in its corresponding scenario, and uses the sample images to adjust the generalized model to obtain a data processing model. In other words, the first client adjusts the generalized model using private data of its own scenario. The data processing model obtained after adjustment has better capacity of processing private data of the first client's corresponding scenario. It can be seen that the data processing data has both generalizability of adapting to various scenarios and stronger adaptability to a certain specific scenario. In other words, the data processing model has both versatility and specificity.

In one implementation, the generalized model includes a low-level model and a high-level model, wherein the low-level model is a partial model close to an input layer, and the high-level model is a partial model close to an output layer. The feature extraction model is the low-level model. For example, the low-level model can include a small number of convolutional layers, BN layers, and pooling layers. The low-level model is configured for performing feature extraction on input data. In step S503 of adjusting the generalized model, only the high-level model instead of the low-level model may be adjusted.

S504: sending the private sample images to the server.

In one implementation, a first client can further receive an encryption model sent by the server; in this implementation, step S504 can include: encrypting the private sample images using the encryption model to obtain encrypted data; and sending the encrypted data to the server.

In this implementation, by encrypting private sample images of the first client before sending them to the server, data safety is increased.

In one implementation, the encryption model includes a feature extraction model; in step S504, the first client can perform feature extraction on the private sample images using the feature extraction model to obtain encrypted data.

In an example, the first client performs feature extraction on private sample images using the feature extraction model to obtain features of the private sample images, and sends the features as encrypted data to the server. That is, in this example, feature extraction is used as a means of encryption. Feature extraction can be understood as lossy processing of sample images, thus achieving an effect of encrypting the sample images and ensuring the safety of the sample

images. Even if an unauthorized user obtains features of the sample images, he/she cannot restore the sample images.

In this example, after receiving the features of the private sample images, the server can directly use the features of the private sample images to adjust the generalized model.

5 In another example, the encryption model includes a feature extraction model and a compression encoding model; in step S504, the first client can perform feature extraction on the private sample images using the feature extraction model to obtain features of the private sample images; perform compression encoding on the features of the private sample images using the compression encoding model to obtain encoded features as encrypted data.

10 In this embodiment, the server can decode the encoded features by means of a decoding model to obtain decoded features, and adjust the generalized model based on the decoded features.

15 In this embodiment, using feature extraction and feature encoding as means of encryption, is equivalent to perform two operations of lossy processing on the sample images. On one hand, it achieves an effect of encrypting the sample images, ensuring the safety of the sample images and making it impossible to restore the sample images even if an unauthorized user obtains encoded features. On the other hand, feature encoding can reduce data quantity, reduce bandwidth consumption during data transfer, and also reduce memory consumption during subsequent data processing.

20 Alternatively, in another implementation, the first client can use another encryption algorithm model to encrypt sample images, and the server uses a corresponding decryption algorithm model to decrypt the encrypted data. The specific encryption and decryption algorithms are not limited.

25 By applying the embodiment shown in Fig. 5 of the present application, in a first aspect, the server obtains the generalized model by training based on encrypted data sent by a plurality of clients, wherein each of the clients corresponds respectively to a different scenario. That is, the server obtains the generalized model based on multi-scenario data. Therefore, the generalized model can adapt to a plurality of scenarios, which increases the generalized model's output accuracy in different scenarios. In a second aspect, the first client adjusts the generalized model

using its own scenario's private sample images to obtain a data processing model. The data processing model can have both generalizability of adapting to a plurality of scenarios, and stronger adaptability to a specific scenario. In other words, the data processing model has both versatility and specificity, which increases the data processing model's output accuracy in different scenarios. In a third aspect, in some implementations, the first client encrypts sample images using an encryption model and then sends encrypted data to the server, instead of directly sending the sample images to the server. Thus, data safety is increased.

Correspondingly to the above-described system embodiments, an embodiment of the present application further provides a neural network model adjustment method, applied to a server having stored therein a generalized model. In reference to Fig. 6, the method includes:

S601: receiving private sample images sent by each of clients.

Each of the clients corresponds respectively to a different scenario, and the private sample images are sample images collected in scenarios respectively corresponding to each of the clients.

S602: adjusting the generalized model stored in the server based on the private sample images sent by each of the clients.

S603: sending the adjusted generalized model to each of the clients,

so that each of the clients adjusts the generalized model using its own private sample images.

In one implementation, an encryption model is further stored in the server; the method further includes: sending the encryption model to each of the clients; in this implementation, step S601 can include: receiving encrypted data sent by each of the clients, wherein the encrypted data is data obtained by the clients encrypting the sample images collected in the clients' own corresponding scenarios using the encryption model.

In this implementation, with the clients encrypting private sample images before sending them to the server, data safety is increased.

In one implementation, the generalized model includes a low-level model and a high-level model, wherein, the low-level model is a partial model close to an input layer, and the high-level

model is a partial model close to an output layer. The encryption model includes a feature extraction model, and the feature extraction model is the low-level model; the encrypted data are data obtained by the clients performing feature extraction on sample images collected in the clients' own corresponding scenarios using the feature extraction model.

5 In this implementation, step S602 can include: inputting the encrypted data into the high-level model, and adjusting the high-level model based on an output result of the high-level model.

10 In another implementation, the generalized model includes a low-level model and a high-level model, wherein the low-level model is a partial model close to an input layer, and the high-level model is a partial model close to an output layer. The encryption model includes a feature extraction model and a compression encoding model, and the feature extraction model is the low-level model. The encrypted data are data obtained by the clients performing, sequentially, feature extraction and compression encoding on sample images collected in the clients' own corresponding scenarios using the feature extraction model and the compression
15 encoding model.

 In this implementation, step S602 can include: decoding the encrypted data by means of the decoding model to obtain decoded features; inputting the decoded features into the high-level model, and adjusting the high-level model based on an output result of the high-level model.

20 By applying the embodiment shown in Fig. 6 of the present application, in a first aspect, the server obtains the generalized model by training based on encrypted data sent by a plurality of clients, and each of the clients corresponds respectively to a different scenario. That is, the server obtains the generalized model based on multi-scenario data. Therefore, the generalized model can adapt to a plurality of scenarios, which increases the generalized model's output accuracy in different scenarios. In a second aspect, the first client adjusts the generalized model using its own
25 scenario's private sample images to obtain a data processing model. The data processing model can have both generalizability of adapting to a plurality of scenarios, and stronger adaptability to a specific scenario. In other words, the data processing model has both versatility and specificity, which increases the data processing model's output accuracy in different scenarios. In a third aspect, in some implementations, the first client encrypts sample images using an encryption
30 model and then sends encrypted data to the server, instead of directly sending the sample images

to the server. Thus, data safety is increased.

An embodiment of the present application further provides an electronic device including a processor 701 and a memory 702, as shown in Fig. 7;

the memory 702 is configured for storing a computer program;

5 the processor 701 is configured for, when executing the program stored in the memory 702, carrying out any above-described neural network model adjustment method.

The memory mentioned in the above electronic device may include a random access memory (RAM), and may also include a non-volatile memory (NVM), such as at least one disk memory. Optionally, the memory may also be at least one storage device located far away from
10 the aforementioned processor.

The above-mentioned processor can be a general-purpose processor, including a central processing unit (CPU), a network processor (NP), etc.; it can also be a digital signal processor (DSP), application specific integrated circuit (ASIC), Field-Programmable Gate Array (FPGA) or other programmable logic devices, discrete gate or transistor logic devices, discrete hardware
15 components.

In yet another embodiment provided by the present application, a computer readable storage medium is further provided, which has stored therein a computer program. The computer program, when executed by a processor, causes the processor to carry out any above-described neural network model adjustment method.

20 In yet another embodiment provided by the present application, a computer program product containing instructions is further provided which, when running on a computer, causes the computer to carry out any above-described neural network model adjustment method.

In the above embodiments, all or part of them may be implemented by software, hardware, firmware or any combination thereof. When implemented using software, it may be implemented
25 in whole or in part in the form of a computer program product. The computer program product includes one or more computer instructions. When the computer program instructions are loaded and executed on the computer, the processes or functions according to the embodiments of the present application will be generated in whole or in part. The computer can be a general purpose

computer, a special purpose computer, a computer network, or other programmable devices. The computer instructions may be stored in a computer-readable storage medium or transmitted from one computer-readable storage medium to another computer-readable storage medium, for example, the computer instructions may be transmitted from a website, computer, server or data center to another website site, computer, server, or data center by wired (eg, coaxial cable, fiber optic, digital subscriber line (DSL)) or wireless (eg, infrared, wireless, microwave, etc.) means. The computer-readable storage medium may be any available medium that can be accessed by a computer, or an integrated data storage device such as a server or a data center containing one or more available media storage. The available medium may be a magnetic medium (for example, a floppy disk, a hard disk, or a magnetic tape), an optical medium (for example, DVD), or a semiconductor medium (for example, a Solid State Disk (SSD)).

It should be noted that in the specification, relational terms such as first and second are only used to distinguish one entity or operation from another entity or operation, and do not necessarily require or imply that there is an actual relationship between these entities or operations. Furthermore, the term "comprise", "include" or any other variation thereof is intended to cover a non-exclusive inclusion such that a process, method, article, or apparatus comprising a set of elements includes not only those elements, but also include other elements not expressly listed or also include elements inherent in such a process, method, article, or device. Without further limitations, an element defined by the phrase "comprising a ..." does not exclude the presence of additional identical elements in the process, method, article or apparatus comprising said element.

Each embodiment in this specification is described in a related manner, the same and similar parts of each embodiment can be referred to each other, and each embodiment focuses on the differences from other embodiments. In particular, for the embodiment of the neural network model adjustment method applied to the first client, the embodiment of the neural network model adjustment method applied to the server, the embodiment of the device, the embodiment of the computer-readable storage medium, and the embodiment of the computer program product, since they are basically similar to the embodiment of the neural network model adjustment system, the description is relatively simple, and for relevant parts, reference can be made to the part of the description of the system embodiment.

The above descriptions are only preferred embodiments of the present application, and are not intended to limit the protection scope of the present application. Any modifications, equivalent replacements, improvements, etc. made within the spirit and principles of the present application are included within the protection scope of this application.

5

10

CLAIMS

What is claimed is:

1. A neural network model adjustment system, comprising a server and a plurality of clients, wherein each of the clients correspond respectively to a different scenario; wherein,

5 the server is configured for sending a generalized model to each of the clients;

the clients are configured for obtaining sample images collected in scenarios respectively corresponding to the clients as private sample images; adjusting the generalized model using the private sample images to obtain a data processing model; and sending the private sample images to the server;

10 the server is further configured for adjusting the generalized model based on the private sample images sent by each of the clients, and sending the adjusted generalized model to each of the clients.

2. The system according to claim 1, wherein,

the server is further configured for sending an encryption model to each of the clients;

15 the clients are specifically configured for encrypting their own private sample images using the encryption model to obtain encrypted data; and sending the encrypted data to the server;

the server is specifically configured for adjusting the generalized model based on the encrypted data sent by each of the clients.

20 3. The system according to claim 2, wherein, the encryption model comprises a feature extraction model;

the clients are specifically configured for performing feature extraction on the private sample images by the feature extraction model to obtain the encrypted data.

4. The system according to claim 3, wherein, the encryption model further comprises a compression encoding model;

25 the clients are specifically configured for performing the feature extraction on the private sample images by the feature extraction model to obtain features of the private sample images; performing compression encoding on the features of the private sample images by the

compression encoding model to obtain encoded features; sending the encoded features to the server as the encrypted data;

the server is specifically configured for decoding the encoded features by a decoding model to obtain decoded features, and adjusting the generalized model based on the decoded features.

5 5. The system according to claim 3 or 4, wherein, the generalized model comprises a low-level model and a high-level model, wherein the low-level model is a partial model close to an input layer, and the high-level model is a partial model close to an output layer; the feature extraction model is the low-level model.

10 6. A neural network model adjustment method, which is applied to a first client, wherein the method comprises:

receiving a generalized model sent by a server, wherein the generalized model is obtained by the server adjusting the generalized model based on encrypted data sent by a plurality of clients, and each of the clients corresponds respectively to a different scenario;

15 obtaining sample images collected in a scenario corresponding to the first client as private sample images;

adjusting the generalized model using the private sample images to obtain a data processing model;

sending the private sample images to the server.

7. The method according to claim 6, wherein, the method further comprises:

20 receiving an encryption model sent by the server;

sending the private sample images to the server comprises:

encrypting the private sample images by the encryption model to obtain encrypted data; and sending the encrypted data to the server.

25 8. The method according to claim 7, wherein, the encryption model comprises a feature extraction model; encrypting the private sample images by the encryption model to obtain the encrypted data, comprises:

performing feature extraction on the private sample images by the feature extraction model

to obtain the encrypted data.

9. The method according to claim 8, wherein, the encryption model further comprises a compression encoding model; performing feature extraction on the private sample images by the feature extraction model to obtain the encrypted data, comprises:

5 performing the feature extraction on the private sample images by the feature extraction model to obtain features of the private sample images;

performing compression encoding on the features of the private sample images by the compression encoding model to obtain encoded features as the encrypted data.

10. The method according to claim 6, wherein, the method further comprises:

10 after receiving the generalized model, determining whether the generalized model satisfies a preset performance condition;

if the generalized model satisfies the preset performance condition, determining the generalized model as the data processing model;

15 if the generalized model does not satisfy the preset performance condition, executing the step of obtaining the sample images collected in the scenario corresponding to the first client as private sample images.

11. A neural network model adjustment method, applied to a server having stored therein a generalized model, wherein the method comprises:

20 receiving private sample images sent by each of clients, wherein each of the clients corresponds respectively to a different scenario, and the private sample images are sample images collected in scenarios respectively corresponding to the clients;

adjusting the generalized model stored in the server based on the private sample images sent by each of the clients;

25 sending the adjusted generalized model to each of the clients, so that each of the clients adjusts the generalized model using its own private sample images.

12. An electronic device, comprising a processor and a memory, wherein

the memory is configured for storing a computer program;

the processor is configured for, when executing the program stored in the memory, carrying out the method according to any one of claims 6 to 11.

13. A computer readable storage medium having stored therein a computer program, wherein the computer program, when executed by a processor, causes the processor to carry out
5 the method according to any one of claims 6 to 11.

14. A computer program product containing instructions which, when running on a computer, causes the computer to carry out the method according to any one of claims 6 to 11.