

IMAGE PROCESSING METHOD AND APPARATUS, ELECTRONIC DEVICE AND STORAGE MEDIUM

CROSS-REFERENCE TO RELATED APPLICATION

[0001] The present disclosure is based on and claims priority to Chinese patent application No. 202010366985.5, filed on April 30, 2020, the disclosure of which is hereby incorporated by reference in its entirety.

TECHNICAL FIELD

[0002] Embodiments of the disclosure relates to the technical field of image processing, and in particular to an image processing method and apparatus, an electronic device and a storage medium.

BACKGROUND

[0003] In recent years, in applications such as face recognition and object detection, a neural network model is usually used to extract image features of a to-be-processed image, so as to perform face recognition, object detection and the like according to the image features of the to-be-processed image. In order to improve the accuracy of face recognition, object detection and other applications, the neural network model usually needs further development, that is, increasing the number of layers and neurons of the neural network model. However, the more layers and neurons of the neural network model are, the greater the calculation amount of the neural network model is. In practice, due to the performance limitations of a processor and a memory and the real-time requirements of the applications, not only the accuracy of the neural network model needs improvement, but also the calculation amount of the neural network model is to be limited.

[0004] In some cases, the number of neurons of the neural network model needs to be reduced by means of model pruning. In the related art, a threshold-based method is usually employed to prune the neurons with smaller influence on accuracy in the neural network model, but for the methods in the related art, it is difficult to improve the accuracy of the pruned neural network model while controlling the calculation amount.

SUMMARY

[0005] The disclosure provides an image processing method and apparatus, an electronic device and a storage medium.

[0006] The embodiments of the disclosure provide an image processing method, which may include that: a to-be-processed image is input to a neural network model; and image features of the to-be-processed image are output through the neural network model. The neural network model may include at least two intermediate layers, and each intermediate layer may include a plurality of neurons. The neural network model is obtained by performing pruning processing on an initial neural network model based on the retention probabilities of the neurons in the initial neural network model, and the retention probabilities of the neurons in the initial neural network model are acquired based on a preset statistical model.

[0007] Thus, since the retention probabilities of the neurons are trained according to the preset statistical model, based on the retention probabilities of the neurons determined by the preset statistical model, the initial neural network model may be pruned under a preset calculation amount requirement, and the acquired neural network model can improve the accuracy of the model.

[0008] In one embodiment, before the to-be-processed image is input to the neural network model, the method may further include that: a to-be-trained neural network model is acquired based on the retention probabilities of the neurons in the initial neural network model; and the to-be-trained neural network model is trained and the neural network model is obtained.

[0009] Thus, the to-be-trained neural network model is acquired based on the retention probabilities of the neurons in the initial neural network model, so that the accuracy of the acquired neural network model can be made higher.

[0010] In one embodiment, the preset statistical model is a Markov model. Before the to-be-trained neural network model is acquired based on the retention probabilities of the neurons in the initial neural network model, the method may further include that: the retention probabilities of the neurons in the initial neural network model are acquired according to the Markov model. The transition probability of the k^{th} neuron in any layer of the Markov model is used to represent the retention probability of the k^{th} neuron on the premise that the first $k-1$ neurons are retained, and k is an integer greater than or equal to 1.

[0011] Thus, in the process of training the Markov model, the probability in the Markov model is defined as a parameter that may be learned, and the Markov model and the initial neural network model are integrated for training to optimize parameters in the Markov model.

[0012] In one embodiment, the Markov model is acquired through N times of iterative training and N is an integer greater than or equal to 1. Herein, M is an integer greater than or equal to 1 and less than or equal to N in the M^{th} iterative training. The method may further include that: pruning processing is performed on a $(M-1)^{\text{th}}$ initial neural network model based on a $(M-1)^{\text{th}}$ Markov model in the M^{th} iteration process, and a subnetwork model of the $(M-1)^{\text{th}}$ initial neural network model is acquired; in the M^{th} iteration process, the subnetwork model of the $(M-1)^{\text{th}}$ initial neural network model is trained and a first loss function is acquired, and a first gradient determined by the first loss function is returned to the neuron corresponding to the update of the $(M-1)^{\text{th}}$ initial neural network model and an M^{th} initial neural network model is obtained, the first loss function being used to represent the accuracy of the subnetwork model of the $(M-1)^{\text{th}}$ initial neural network model; in the M^{th} iteration process, a sample image is input to the M^{th} initial neural network model, and an output result of each neuron of the M^{th} initial neural network model is multiplied by the retention probability of the corresponding neuron in the $(M-1)^{\text{th}}$ Markov model; in the M^{th} iteration process, a second loss function

of the M^{th} initial neural network model is acquired, the second loss function being used to represent the accuracy and/or calculation amount of the M^{th} initial neural network model; and in the M^{th} iteration process, a second gradient determined according to the second loss function is returned to the $(M-1)^{\text{th}}$ Markov model, and an M^{th} Markov model is obtained.

[0013] Thus, in the process of training the Markov model, the probability in the Markov model is defined as a parameter that may be learned, and the Markov model and the initial neural network model are integrated for training to optimize parameters in the Markov model. The method provided by the embodiments of the disclosure is high in training efficiency, and the neural network model acquired after pruning the initial neural network model by adopting the trained Markov model may meet the preset calculation amount requirement and improve the accuracy of the model at the same time.

[0014] In one embodiment, before the to-be-trained neural network model is trained and the neural network model is obtained, the method may further include: randomizing processing is performed on the parameters of the neurons in the to-be-trained neural network model.

[0015] Thus, randomizing processing is performed on the parameters of the neurons in the to-be-trained neural network model, so that the interference of the parameters of the neurons in the initial neural network model to the parameters of the neurons in the to-be-trained neural network model acquired after pruning may be reduced.

[0016] In one embodiment, the neural network model is obtained by performing pruning processing on the initial neural network model based on the retention probabilities of the neurons in any intermediate layer or the retention probabilities of the neurons in multiple intermediate layers in the initial neural network model.

[0017] Thus, not only pruning processing may be performed on the initial neural network model based on the retention probabilities of the neurons in any intermediate layer, but also pruning processing may be performed on the initial neural network model based on the retention probabilities of the neurons in multiple intermediate layers, thus improving the diversity of the pruning processing mode.

[0018] In one embodiment, after the image features of the to-be-processed image are output through the neural network model, the method may further include: at least one of the following processes is performed according to the obtained image features: pedestrian retrieval, face recognition, face location detection, face key point detection, human body location detection, human body motion detection, human body key point detection, gesture recognition, hand location detection, pedestrian recognition, vehicle recognition, scenario recognition and in-vivo detection.

[0019] Thus, the image features may be processed in different modes to meet the needs of different application scenarios, and the adaptability of the image feature processing modes is improved.

[0020] In one embodiment, the method may further include: the pruning process of each intermediate layer in the initial neural network model is defined as one Markov model; pruning processing is performed on the initial neural network model according to the Markov model and a pruned neural network model is obtained; randomizing processing is performed on parameters of neurons in the pruned neural network model and a to-be-trained neural network model is obtained; and the to-be-trained neural network model is trained using a target data set and the neural network model is obtained.

[0021] In one embodiment, the Markov model is acquired through N times of iterative training and N is an integer greater than or equal to 1. Herein, M is an integer greater than or equal to 1 and less than or equal to N in the M^{th} iterative training. The operation of performing pruning processing on the initial neural network model and obtaining the pruned neural network model may include: the retention probabilities of the neurons in the $(M-1)^{\text{th}}$ initial neural network model are acquired according to the $(M-1)^{\text{th}}$ Markov model; the subnetwork model of the $(M-1)^{\text{th}}$ Markov model is sampled in the $(M-1)^{\text{th}}$ initial neural network model according to the retention probabilities of the neurons in the $(M-1)^{\text{th}}$ initial neural network model; the $(M-1)^{\text{th}}$ initial neural network model is updated using the first gradient returned by the trained subnetwork model of the $(M-1)^{\text{th}}$ Markov model and the M^{th} initial neural network model is obtained; the second gradient is determined according to output results of the neurons in the M^{th} initial neural network model and the retention probabilities of the neurons in the $(M-1)^{\text{th}}$ initial neural network

model in the case that the sample image is input to the M^{th} initial neural network model; and the $(M-1)^{\text{th}}$ Markov model is optimized using the second gradient and the M^{th} Markov model is obtained.

[0022] The embodiments of the disclosure provide an image processing apparatus, which may include: an input module, configured to input a to-be-processed image to a neural network model; and an output module, configured to output image features of the to-be-processed image through the neural network model. The neural network model may include at least two intermediate layers, and each intermediate layer may include a plurality of neurons. The neural network model is obtained by performing pruning processing on an initial neural network model based on the retention probabilities of the neurons in the initial neural network model, and the retention probabilities of the neurons in the initial neural network model are acquired based on a preset statistical model.

[0023] In one embodiment, the apparatus may further include a processing module, configured to acquire a to-be-trained neural network model based on the retention probabilities of the neurons in the initial neural network model before the input module inputs the to-be-processed image to the neural network model; and train the to-be-trained neural network model and obtain the neural network model.

[0024] In one embodiment, the preset statistical model is a Markov model. The processing module is further configured to, before acquiring the to-be-trained neural network model based on the retention probabilities of the neurons in the initial neural network model, acquire the retention probabilities of the neurons in the initial neural network model according to the Markov model. The transition probability of the k^{th} neuron in any layer of the Markov model is used to represent the retention probability of the k^{th} neuron on the premise that the first $k-1$ neurons are retained, k being an integer greater than or equal to 1.

[0025] In one embodiment, the Markov model is acquired through N times of iterative training and N is an integer greater than or equal to 1. Herein, M is an integer greater than or equal to 1 and less than or equal to N in the M^{th} iterative training. The processing module is further configured to perform, in the M^{th} iteration process, pruning processing on a $(M-1)^{\text{th}}$ initial neural network model based on a $(M-1)^{\text{th}}$ Markov model, and acquire

a subnetwork model of the $(M-1)^{\text{th}}$ initial neural network model; train, in the M^{th} iteration process, the subnetwork model of the $(M-1)^{\text{th}}$ initial neural network model and acquire a first loss function, and return a first gradient determined by the first loss function to the neuron corresponding to the update of the $(M-1)^{\text{th}}$ initial neural network model and obtain an M^{th} initial neural network model; input, in the M^{th} iteration process, a sample image to the M^{th} initial neural network model, and multiply an output result of each neuron of the M^{th} initial neural network model by the retention probability of the corresponding neuron in the $(M-1)^{\text{th}}$ Markov model; acquire, in the M^{th} iteration process, a second loss function of the M^{th} initial neural network model; and return, in the M^{th} iteration process, a second gradient determined according to the second loss function to the $(M-1)^{\text{th}}$ Markov model and obtain an M^{th} Markov model. The first loss function is used to represent the accuracy of the subnetwork model of the $(M-1)^{\text{th}}$ initial neural network model, and the second loss function is used to represent the accuracy and/or calculation amount of the M^{th} initial neural network model.

[0026] In one embodiment, the processing module is further configured to, before training the to-be-trained neural network model and obtaining the neural network model, perform randomizing processing on parameters of the neurons in the to-be-trained neural network model.

[0027] In one embodiment, the neural network model is obtained by performing pruning processing on the initial neural network model based on the retention probabilities of the neurons in any intermediate layer or the retention probabilities of the neurons in multiple intermediate layers in the initial neural network model.

[0028] In one embodiment, the processing module is further configured to, after the output module outputs the image features of the to-be-processed image through the neural network model, perform at least one of the following processes according to the obtained image features: pedestrian retrieval, face recognition, face location detection, face key point detection, human body location detection, human body motion detection, human body key point detection, gesture recognition, hand location detection, pedestrian recognition, vehicle recognition, scenario recognition and in-vivo detection.

[0029] In one embodiment, the apparatus may further include: a defining module,

configured to define the pruning process of each intermediate layer in the initial neural network model as one Markov model; a pruning module, configured to perform pruning processing on the initial neural network model according to the Markov model and obtain a pruned neural network model; a randomizing module, configured to perform randomizing processing on the parameters of the neurons in the pruned neural network model and obtain a to-be-trained neural network model; and a training module, configured to train the to-be-trained neural network model using a target data set and obtain the neural network model.

[0030] In one embodiment, the Markov model is acquired through N times of iterative training and N is an integer greater than or equal to 1. Herein, M is an integer greater than or equal to 1 and less than or equal to N in the M^{th} iterative training. The pruning module may include: an acquisition unit, configured to acquire the retention probabilities of the neurons in the $(M-1)^{\text{th}}$ initial neural network model according to the $(M-1)^{\text{th}}$ Markov model; a sampling unit, configured to sample the subnetwork model of the $(M-1)^{\text{th}}$ Markov model in the $(M-1)^{\text{th}}$ initial neural network model according to the retention probabilities of the neurons in the $(M-1)^{\text{th}}$ initial neural network model; an update unit, configured to update the $(M-1)^{\text{th}}$ initial neural network model using the first gradient returned by the trained subnetwork model of the $(M-1)^{\text{th}}$ Markov model and obtain the M^{th} initial neural network model; a determination unit, configured to determine the second gradient according to the output results of the neurons in the M^{th} initial neural network model and the retention probabilities of the neurons in the $(M-1)^{\text{th}}$ initial neural network model in the case that the sample image is input to the M^{th} initial neural network model; and an optimization unit, configured to optimize the $(M-1)^{\text{th}}$ Markov model using the second gradient and obtain the M^{th} Markov model.

[0031] The embodiments of the disclosure provide an electronic device, which may include: at least one processor and a memory. The memory stores a computer execution instruction. The at least one processor executes the computer execution instruction stored in the memory to enable the apparatus to execute any of the described methods.

[0032] The embodiments of the disclosure provide a computer readable storage medium. A computer execution instruction is stored on the computer readable storage

medium. Any of the described methods is implemented in the case that the computer execution instruction is executed by a processor.

[0033] The embodiments of the disclosure provide a computer program product. The computer program product may include a non-transient computer readable storage medium storing a computer program. The computer program is operable to cause a computer to execute some or all of the steps described in the image processing method of the embodiments of the disclosure. The computer program product may be a software package.

[0034] It is to be understood that the above general descriptions and the following detailed descriptions are only exemplary and explanatory, and do not limit the embodiments of the disclosure.

[0035] Other characteristics and aspects of the disclosure will become apparent from the following detailed descriptions of the exemplary embodiments with reference to the accompanying drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

[0036] In order to describe the technical solutions in the embodiments of the disclosure or in the related art more clearly, the drawings required to be used in descriptions about the embodiments or the related art will be simply introduced below, obviously, the drawings described below are some embodiments of the disclosure, and other drawings may further be obtained by those of ordinary skill in the art according to the drawings without creative work.

[0037] FIG. 1 is a schematic diagram of an application scenario of an image processing method according to an embodiment of the disclosure.

[0038] FIG. 2 is a flowchart of an image processing method according to an embodiment of the disclosure.

[0039] FIG. 3 is a schematic diagram of an intermediate layer according to an

embodiment of the disclosure.

[0040] FIG. 4 is a schematic diagram of acquisition of a neural network model based on an initial neural network model according to an embodiment of the disclosure.

[0041] FIG. 5 is a schematic diagram of transition probability of any layer of a Markov model according to an embodiment of the disclosure.

[0042] FIG. 6 is a flowchart of the M^{th} iterative training of a Markov model according to an embodiment of the disclosure.

[0043] FIG. 7 is a structural schematic diagram of an image processing apparatus according to an embodiment of the disclosure.

[0044] FIG. 8 is a structural schematic diagram of an electronic device according to an embodiment of the disclosure.

[0045] Specific embodiments of the disclosure have been shown by the above-mentioned drawings, and more detailed descriptions will be made later. These drawings and written descriptions are not intended to limit the scope of the concept of the disclosure in any way, but to explain the concept of the disclosure to those skilled in the art with reference to specific embodiments.

DETAILED DESCRIPTION

[0046] In order to make the purposes, technical solutions and advantages of the embodiments of the disclosure clearer, the technical solutions in the embodiments of the disclosure will be clearly and completely described below in combination with the drawings in the embodiments of the disclosure, and it is apparent that the described embodiments are only a part rather all of embodiments of the disclosure. All other embodiments obtained by those of ordinary skill in the art based on the embodiments of the disclosure without creative efforts shall fall within the protection scope of the disclosure.

[0047] The embodiments of the disclosure may be applied to an electronic device such as a terminal device, a computer system, or a server, which may operate with numerous other general-purpose or dedicated computing system environments or configurations. Examples of well-known terminal devices, computing systems, environments and/or configurations suitable for use with the electronic device such as the terminal device, the computer system or the server include, but are not limited to, a personal computer system, a server computer system, a thin client, a thick client, a hand-held or laptop device, a microprocessor-based system, a Set Top Box (STB), a programmable consumer electronic, a network personal computer, a small computer system, a large computer system and a distributed cloud computing technology environment including any of the above systems.

[0048] The electronic device such as the terminal device, the computer system, or the server may be described in the general context of computer system executable instructions (such as a program module) executed by the computer system. Generally, the program module may include a routine, a program, an object program, a component, logic, a data structure, etc., which executes a specific task or implements a specific abstract data type. The computer system/server may be implemented in a distributed cloud computing environment, and in the distributed cloud computing environment, the task is executed by a remote processing device linked through a communication network. In the distributed cloud computing environment, the program module may be located on a local or remote computing system storage medium including a storage device.

[0049] In the following embodiments of the disclosure, the execution body of the method embodiments are explained by taking the terminal device as an example.

[0050] The technical solutions of image processing in the disclosure may generally be applied to applications that need to extract image features from images. For example, the technical solutions of image processing in the disclosure may be applied to applications such as pedestrian retrieval, a face recognition result, face location detection, face key point detection, human body location detection, human body motion detection, human body key point detection, gesture recognition, hand location detection, pedestrian recognition, vehicle recognition, scenario recognition and living body detection. It is to

be seen from the above that various applications such as pedestrian retrieval, face recognition, face location detection, face key point detection, human body location detection, human body motion detection, human body key point detection, gesture recognition, hand location detection, pedestrian recognition, vehicle recognition, scenario recognition and living body detection may all use the neural network model formed based on pruning in the disclosure. The disclosure does not limit the technical solutions of image processing and the application scenarios of the neural network model formed based on pruning.

[0051] In order to better understand the application scenario in the disclosure, exemplarily, FIG. 1 is a schematic diagram of an application scenario of an image processing method according to an embodiment of the disclosure. As shown in FIG. 1, a terminal device 11 and a camera 12 are included. The terminal device 11 may collect a face image in front of the camera through the camera 12. The terminal device 11 may process the collected image based on the image processing method in the disclosure to recognize the face in the image and execute subsequent operations. For example, the terminal device and the like may be unlocked based on the collected face information.

[0052] According to the image processing method and apparatus, the electronic device and the storage medium according to the embodiments of the disclosure, in the method, the neural network model configured to acquire the image features is obtained by performing pruning processing on an initial neural network model based on the retention probabilities of neurons in the initial neural network model, and the retention probability is determined by a statistical model. In the disclosure, based on the method, the initial neural network model may be pruned under a preset calculation requirement, and the acquired neural network model can also improve the accuracy of the model.

[0053] FIG. 2 is a flowchart of an image processing method according to an embodiment of the disclosure. As shown in FIG. 2, the method may include the following operations.

[0054] At S101, a to-be-processed image is input to a neural network model.

[0055] The to-be-processed image may be a picture or photo, or may be a image frame

in a video.

[0056] The contents contained in the to-be-processed image may be determined according to an actual application scenario. For example, in the case that the image processing method in the disclosure is applied to an application related to a human face, such as pedestrian retrieval, face recognition, face location detection or face key point detection, the to-be-processed image may be an image containing the human face, and the human face in the to-be-processed image may be a front face or a side face. For example, in the case that the image processing method in the disclosure is applied to an application related to a human body, such as pedestrian retrieval, pedestrian recognition, human body key point detection, human body location detection or human body motion detection, the to-be-processed image may be an image containing the human body. For example, in the case that the image processing method in the disclosure is applied to an application related to the hands, such as gesture detection or hand location detection, the to-be-processed image may be an image containing the hands. For example, in the case that the image processing method in the disclosure is applied to an application related to a vehicle, such as vehicle recognition, the to-be-processed image may be an image containing the vehicle. For example, in the case that the image processing method in the disclosure is applied to an application related to a scenario, such as scenario recognition, the to-be-processed image may be an image containing the scenario. For example, in the case that the image processing method in the disclosure is applied to an application related to a living body, such as living body detection, the to-be-processed image may be an image containing the living body (such as a living person, a living cat, a living dog, etc.).

[0057] The neural network model includes at least two intermediate layers, and each intermediate layer includes a plurality of neurons. FIG. 3 is a schematic diagram of an intermediate layer according to an embodiment of the disclosure. As shown in FIG. 3, the intermediate layer is located between an input layer and an output layer. The intermediate layer may be, for example, a hidden layer. For example, in the case that the neural network model is a convolutional neural network model, the intermediate layer may be a convolutional layer.

- [0058] The neural network model may be any one of the following neural network models: Residual Network (ResNet), MobileNet, and MobileNet-v2.
- [0059] The neural network model may be obtained by performing pruning processing on the initial neural network model based on the retention probabilities of the neurons in the initial neural network model, and the retention probabilities of the neurons in the initial neural network model are acquired based on a preset statistical model.
- [0060] In a possible implementation mode, the neural network model is obtained by performing pruning processing on the initial neural network model based on the retention probabilities of the neurons in any intermediate layer in the initial neural network model. That is, only the neurons in a certain intermediate layer in the initial neural network model may be pruned.
- [0061] In another possible implementation mode, the neural network model is obtained by performing pruning processing on the initial neural network model based on the retention probabilities of the neurons in multiple intermediate layers in the initial neural network model. That is, the neurons of more than or equal to two partial intermediate layers in the initial neural network model may be pruned. Herein, the more than or equal to two partial intermediate layers may be continuous or discontinuous. The neurons in each intermediate layer in the initial neural network model may also be pruned.
- [0062] In the initial neural network model, in the case that a certain neuron is determined to be retained based on the retention probability of the neuron, the neuron is retained in the above neural network model. In the case that it is determined that a certain neuron is not retained based on the retention probability of the neuron, the neuron will not be retained in the above neural network model, that is, the neuron will be pruned. Since the retention probability of the neuron is trained according to the preset statistical model, based on the retention probability of the neuron determined by the preset statistical model, the initial neural network model may be pruned under a preset calculation amount requirement, and the acquired neural network model can improve the accuracy of the model.

- [0063] At S102, image features of the to-be-processed image are output through the neural network model.
- [0064] After the to-be-processed image is processed by the neurons in each layer of the neural network model, the image features of the to-be-processed image are output, and the image features may be, for example, feature images of the to-be-processed image.
- [0065] In an implementation mode, after S102, the method may further include the following operations.
- [0066] At least one of the following is performed according to the obtained image features: pedestrian retrieval, face recognition, face location detection, face key point detection, human body location detection, human body motion detection, human body key point detection, gesture recognition, hand location detection, pedestrian recognition, vehicle recognition, scenario recognition and living body detection. The processing mode depends on the needs of the application scenario.
- [0067] According to the image processing method according to the embodiments of the disclosure, the image features of the to-be-processed image are acquired by inputting the to-be-processed image to the neural network model. The neural network model is obtained by performing pruning processing on the initial neural network model based on the retention probabilities of the neurons in the initial neural network model. Since the retention probabilities of the neurons are trained according to the preset statistical model, based on the retention probabilities of the neurons determined by the preset statistical model, the initial neural network model may be pruned under the preset calculation amount requirement, and the acquired neural network model can improve the accuracy of the model.
- [0068] The method of acquiring the neural network model according to the initial neural network model is implemented as follows.
- [0069] In a possible implementation mode, before S101, the neural network model may be acquired based on the retention probabilities of the neurons in the initial neural network model.

[0070] In another possible implementation mode, before S101, a to-be-trained neural network model may be acquired based on the retention probabilities of the neurons in the initial neural network model; and the to-be-trained neural network model is trained and the neural network model is obtained.

[0071] Exemplarily, the to-be-trained neural network model may be trained for a preset number of times using the sample data set, and then the neural network model is obtained.

[0072] In a possible implementation mode, before training the to-be-trained neural network model and obtaining the neural network model, randomizing processing may further be performed on parameters of the neurons in the to-be-trained neural network model. Randomizing processing is performed on the parameters of the neurons in the to-be-trained neural network model, so that the interference of the parameters of the neurons in the initial neural network model to the parameters of the neurons in the to-be-trained neural network model subjected to the pruning may be reduced.

[0073] Exemplarily, FIG. 4 is a schematic diagram of acquisition of a neural network model based on an initial neural network model according to an embodiment of the disclosure. As shown in FIG. 4, it is assumed that the initial neural network model includes 2 intermediate layers (the input layer and the output layer are not shown), each intermediate layer includes 10 neurons, and initial parameters of the neurons in the initial neural network model are random. Pruning is performed on the initial neural network model based on the retention probabilities of the neurons in the initial neural network model. The acquired to-be-trained neural network model includes, for example, two intermediate layers. The first intermediate layer retains the first six neurons in the first layer of the initial neural network model, and the second intermediate layer retains the first eight neurons in the second layer of the initial neural network model. The to-be-trained neural network model is trained using the sample data set, so that the above-mentioned neural network model may be obtained.

[0074] In the above embodiments, the preset statistical model may be, for example, a Markov model, a Bernoulli distribution model, and the like. In the embodiment, the Markov model is taken as an example for description.

[0075] In the case that the preset statistical model is the Markov model, in the above embodiment, before the to-be-trained neural network model is acquired based on the retention probabilities of the neurons in the initial neural network model, the following steps may also be included.

[0076] According to the Markov model, the retention probabilities P_k of the neurons in the initial neural network model are acquired. The transition probability of the k^{th} neuron in any layer of the Markov model is used to represent the retention probability of the k^{th} neuron given that the first $k-1$ neurons are retained, k being an integer greater than or equal to 1.

[0077] Exemplarily, FIG. 5 is a schematic diagram of transition probability of any layer of a Markov model. As shown in FIG. 5, S_1 is the first retained neuron. S_k is the first k retained neurons. S_{Cout} is the pruned neurons. T indicates the end of a pruning process. P_1 is the probability that the first neuron is retained. $1 - P_1$ is the probability that the first neuron is pruned. P_k is the probability that the k^{th} neuron is retained given that the first $k-1$ neurons are retained. $1 - P_k$ is the probability that the k^{th} neuron is pruned given that the first $k-1$ neurons are retained. P_k may be expressed by formula (1) as follows.

$$[0078] \quad p_k = \begin{cases} 1 & k = 1 \\ \text{sigmoid}(\alpha_k) = \frac{1}{1+e^{-\alpha_k}} & k = 2, \dots, C_{\text{out}}, \alpha_k \in A \end{cases} \quad (1)$$

[0079] where α_k is a trained parameter, and $\text{sigmoid}(\alpha_k)$ is the value of P_k when k is greater than or equal to 2.

[0080] The following embodiments will focus on how to train the Markov model.

[0081] The Markov model is acquired through N iterative trainings and the M^{th} training of the N iterative trainings may be recorded as the M^{th} iterative training. Herein, N is an integer greater than or equal to 1, and the size of N may be determined according to an actual situation. M is an integer greater than or equal to 1 and less than or equal to N .

[0082] FIG. 6 is a flowchart of the M^{th} iterative training of a Markov model according

to an embodiment of the disclosure. As shown in FIG. 6, the method may include the following operations.

[0083] At S201, in the M^{th} iteration process, pruning processing is performed on a $(M-1)^{\text{th}}$ initial neural network model based on a $(M-1)^{\text{th}}$ Markov model, and a subnetwork model of the $(M-1)^{\text{th}}$ initial neural network model is acquired.

[0084] For example, in the first iteration process, a 0^{th} initial neural network model may be, for example, an initial neural network model with random parameters, and a 0^{th} Markov model may be a Markov model with random parameters. The initial neural network model with the random parameters is subjected to pruning processing based on the Markov model with the random parameters, and the subnetwork model of the 0^{th} initial neural network model is acquired.

[0085] Exemplarily, it is assumed that the initial neural network model with the random parameters includes two intermediate layers, and each intermediate layer includes ten neurons. In the first iteration process, after pruning, the first intermediate layer retains the first nine neurons and the second intermediate layer retains the first eight neurons in the subnetwork model of the 0^{th} initial neural network model.

[0086] For another example, in the second iteration process, pruning processing is performed on a first initial neural network model based on a first Markov model, and a subnetwork model of the first initial neural network model is acquired.

[0087] Exemplarily, the first initial neural network model includes two intermediate layers, and each intermediate layer includes ten neurons. In the second iteration process, after pruning, the first intermediate layer retains the first eight neurons and the second intermediate layer retains the first seven neurons in the subnetwork model of the first initial neural network model.

[0088] The subsequent iteration processes are similar to the second iteration process, which will not be described again in the embodiment.

[0089] At S202, in the M^{th} iteration process, the subnetwork model of the $(M-1)^{\text{th}}$ initial neural network model is trained and a first loss function is acquired; and a first

gradient determined by the first loss function is returned to neurons corresponding to the update of the $(M-1)^{\text{th}}$ initial neural network model, and an M^{th} initial neural network model is obtained.

[0090] The first loss function is used to represent the accuracy of the subnetwork model of the $(M-1)^{\text{th}}$ initial neural network model. Exemplarily, the first loss function may be any one of: a loss function of mean-square error (MSE), a loss function of Cross Entropy, and the like.

[0091] Exemplarily, the sample data set may be input into the subnetwork model of the $(M-1)^{\text{th}}$ initial neural network model, the subnetwork model of the $(M-1)^{\text{th}}$ initial neural network model is trained, and , the corresponding first loss function is trained after the training is completed. The first gradient is acquired after the first loss function is derived, and the first gradient is returned to the neuron corresponding to the $(M-1)^{\text{th}}$ initial neural network model by a gradient descent method, and the M^{th} initial neural network model is obtained. The M^{th} initial neural network model may be applied to the $(M+1)^{\text{th}}$ iteration process.

[0092] Taking the second iteration process as an example for descriptions, the first intermediate layer retains the first eight neurons and the second intermediate layer retains the first seven neurons in the subnetwork model of the first initial neural network model. The operation that the subnetwork model of the first initial neural network model is trained and the first loss function is acquired, and the first gradient determined by the first loss function is returned to the neurons corresponding to the first initial neural network model, and a second initial neural network model is obtained after the first gradient is used to update the parameters of the first eight neurons in the first intermediate layer of the first initial neural network model and update the parameters of the first seven neurons in the second intermediate layer.

[0093] At S203, in the M^{th} iteration process, a sample image is input to the M^{th} initial neural network model, and an output result of each neuron of the M^{th} initial neural network model is multiplied by the retention probability of the corresponding neuron in the $(M-1)^{\text{th}}$ Markov model.

[0094] Taking the second iteration process as an example for description, the sample image is input to the second initial neural network model. For any layer of the second initial neural network model, assuming that the output result of the k^{th} neuron in the layer is denoted as O_k , the output result O_k is multiplied by the retention probability of the k^{th} neuron in the layer, which is shown in formula (2).

$$[0095] \quad \widehat{O}_k = O_k \times p(w_k) \quad (2)$$

[0096] $p(w_k)$ is the retention probability of the k^{th} neuron in the layer. $\widehat{O}_k$ is the output result obtained by multiplying the output result O_k of the k^{th} neuron in the layer by the retention probability of the k^{th} neuron in the layer.

[0097] The retaining probability $p(w_k)$ of the k^{th} neuron in any layer may be determined by the transition probability P_k of the k^{th} neuron in any layer, as shown in formula (3) as follows.

$$\begin{aligned} p(w_k) &= p(w_k|w_{k-1})p(w_{k-1}) + p(w_k|\neg w_{k-1})p(\neg w_{k-1}) = p(w_k|w_{k-1})p(w_{k-1}) + 0 \\ &= p(w_1) \prod_{i=2}^k p(w_i|w_{i-1}) = \prod_{i=1}^k p_i \quad (3) \end{aligned}$$

[0098] where $p(w_{k-1})$ is the retention probability of the $(k-1)^{\text{th}}$ neuron in the layer. $p(w_k|w_{k-1})$ is the retention probability of the k^{th} neuron in the case that the $(k-1)^{\text{th}}$ neuron in the layer is retained. $p(\neg w_{k-1})$ is the probability that the $(k-1)^{\text{th}}$ neuron in the layer is not retained. $p(w_k|\neg w_{k-1})$ is the retention probability of the k^{th} neuron in the case that the $(k-1)^{\text{th}}$ neuron in the layer is not retained. The formula (3) is simplified so that the conversion relationship between the retention probability $p(w_k)$ of the k^{th} neuron in any layer and the transition probability P_k of the k^{th} neuron in the layer may be obtained.

[0099] At S204, in the M^{th} iteration process, a second loss function of the M^{th} initial neural network model is acquired.

[00100] The second loss function is used to represent the accuracy and/or calculation amount of the M^{th} initial neural network model.

[00101] Exemplarily, in the case that the second loss function is used to represent the accuracy and calculation amount of the M^{th} initial neural network model, the second loss function may be expressed as formula (4).

$$[00102] \quad \text{LOSS} = \text{loss}_{cls} + \lambda_{reg} \text{loss}_{reg} \quad (4)$$

[00103] where Loss is the second loss function, loss_{cls} is the loss function of the accuracy, loss_{reg} is the loss function of the calculation amount, and λ_{reg} is a preset parameter, so that loss_{cls} and loss_{reg} are in the same order of magnitude.

[00104] Taking the second iteration process as an example for description, the second loss function corresponding to the second initial neural network model may be acquired based on the input sample image and a final output result of the second initial neural network model.

[00105] At S205, in the M^{th} iteration process, the second gradient determined according to the second loss function is returned to the $(M-1)^{\text{th}}$ Markov model, and an M^{th} Markov model is obtained.

[00106] A corresponding second gradient may be acquired after the second loss function is derived, and the second gradient may be expressed by formulas (5) to (7) as follows.

$$[00107] \quad \frac{\partial \text{Loss}}{\partial \alpha_j^{(i)}} = \sum_{k=1}^{C_{\text{out}}} \frac{\partial \text{Loss}}{\partial O_k^{(i)}} \times \frac{\partial O_k^{(i)}}{\alpha_j^{(i)}} \quad (5)$$

$$[00108] \quad \frac{\partial O_k^{(i)}}{\alpha_j^{(i)}} = \begin{cases} 0 & , k < j \\ \frac{\partial p_k}{\partial \alpha_j} O_k^{(i)} \prod_{r \in \{r | r \neq j \text{ and } r \leq k\}} p_r & , k \geq j \end{cases} \quad (6)$$

$$[00109] \quad \frac{\partial p_k}{\partial \alpha_j} = (1 - p_k) p_k \quad (7)$$

[00110] $\alpha_j^{(i)}$ in the formulas (5) to (7) indicates α_k in an i^{th} layer in the formula (1).

$\frac{\partial \text{Loss}}{\partial \alpha_j^{(i)}}$ is the partial derivative of the parameters in the Markov model in the formula (1)

calculated by the second loss function, namely the second gradient. $\frac{\partial \text{Loss}}{\partial O_k^{(i)}}$ is the partial derivative of the output in the formula (3) calculated by the second loss function. $\frac{\partial O_k^{(i)}}{\alpha_j^{(i)}}$ is the partial derivative of the parameters in the Markov model calculated by the output in the formula (3). $O_k^{(i)}$ is the output of the k^{th} neuron in the i^{th} layer in the formula (2).

[00111] Taking the second iteration process as an example for descriptions, the second gradient determined according to the second loss function is returned to the first Markov model, and a second Markov model is obtained.

[00112] In the N^{th} iteration process, the second gradient determined according to the second loss function is returned to a $(N-1)^{\text{th}}$ Markov model, and after the training of the obtained N^{th} Markov model is completed, the N^{th} Markov model is the above-mentioned Markov model.

[00113] In the embodiment of the disclosure, during training of the Markov model, the probability in the Markov model is defined as a parameter that may be learned, and the Markov model and the initial neural network model are integrated for training to optimize the parameters in the Markov model. The method according to the embodiments of the disclosure is high in training efficiency, and the neural network model acquired after pruning the initial neural network model by adopting the trained Markov model may meet the preset calculation amount requirement and improve the accuracy of the model at the same time.

[00114] Deep learning plays a vital role in Artificial Intelligence (AI)-related products, such as object detection and face recognition. The characteristic of deep learning is that depth models with a large number of parameters and a large calculation amount are required. In practical application scenarios, these depth models will have limitations such as hardware computing power and storage space, and meanwhile, a target task will also have requirements on running speed.

[00115] Model pruning is one of the research directions to alleviate the oversize depth model. However, model pruning will lead to the loss of model accuracy. Although

pruning the number of channels of the model can reduce the calculation amount of the model and reduce the storage space, it will reduce the accuracy of the model, thus failing to meet the accuracy requirements of products.

[00116] Model pruning plays a very important role in many scenarios. For example, in pedestrian detection in security, pruning may reduce the storage space and speed up the running speed. In the task of face detection and recognition, model pruning may prune redundant parameters in a large model, thus lightening the model.

[00117] The related art mainly includes performing pruning based on some artificially set rules (such as the size of parameter norm). The model pruning algorithm still has a large loss of model accuracy. The technical efficiency of model search is low, and a reinforcement learning or genetic algorithm is widely used.

[00118] The embodiment of the disclosure defines model pruning as model search, and proposes an image processing method for directly searching the number of channels of the model. After the model is pruned by this method, the calculation amount of the model is significantly reduced, and meanwhile, the accuracy remains unchanged or only slightly decreases while meeting the performance requirements. The following steps S301 to S303 are mainly included.

[00119] At S301, the pruning process of each convolutional layer in an original model is defined as a Markov model.

[00120] The original model may be the initial neural network model, and the initial neural network model may be a Residual Network (ResNet), MobileNet, MobileNet-v2, etc. The initial neural network model may include at least two intermediate layers, and each intermediate layer may include multiple neurons. In the case that the initial neural network model is a convolutional neural network model, the intermediate layer may be a convolutional layer. The retention probabilities of the neurons in each convolutional layer in the original model may be acquired through the Markov model.

[00121] FIG. 5 is a Markov model of a pruning process of a convolutional layer. Referring to FIG. 5, S_k indicates that the layer will retain the first k channels, S_{Cout} indicates pruned neurons, and T indicates the end of the pruning process. P_k indicates the

probability of retaining a k^{th} channel given that the first $k-1$ channels are retained. $1 - P_k$ is the probability of pruning the k^{th} neuron given that the first $k-1$ neurons are retained. P_k is expressed by formula (1) as follows.

$$[00122] \quad P_k = \begin{cases} 1 & k = 1 \\ \text{sigmoid}(\alpha_k) = \frac{1}{1+e^{-\alpha_k}} & k = 2, \dots, C_{\text{out}}, \alpha_k \in A \end{cases} \quad (1)$$

[00123] Channels may be understood as neurons, and redundant channels will be pruned. α_k is a learnable and trained parameter. $\text{sigmoid}(\alpha_k)$ is the value of p_k when k is greater than or equal to 2.

[00124] At S302, pruning processing is performed on the original model.

[00125] Pruning processing may be performed on the original model based on the retention probabilities of the neurons in each convolutional layer. In addition, pruning processing may be performed on the original model only based on the retention probabilities of the neurons in any convolutional layer in the original model. Furthermore, pruning processing may be performed on the original model based on the retention probabilities of the neurons in multiple convolutional layers in the original model. Step S302 may include the following steps S3021 and S3022.

[00126] At S3021, substructures in the original model are sampled using the constructed Markov model, and these substructures are trained on a target data set. Since model parameters of the substructures are shared with the original model, gradients returned from the substructures will be directly updated to the original model.

[00127] The substructure may be understood as the subnetwork model of the initial neural network model, the target data set may be the sample data set, and the gradient may be the first gradient. Assuming that the original model is the first initial neural network model and the Markov model is the first Markov model, pruning processing may be performed on the first initial neural network model based on the first Markov model, and the subnetwork model of the first initial neural network model is sampled. The sample data set is input to the subnetwork model of the first initial neural network model, the subnetwork model of the first initial neural network model is trained, and the corresponding first loss function after training is completed is obtained, the first gradient

is acquired after the first loss function is derived, and the first gradient is returned to the neurons corresponding to the first initial neural network model by the gradient descent method so as to update the first initial neural network model and obtain the second initial neural network model.

[00128] At S3022, the parameters in the Markov model are calculated according to formula (3) as follows the retention probability of each channel is obtained. The retention probability of each neuron is merged with the output result of the corresponding neuron in the original model, and a merged result is $\widehat{O}_k$. The merging mode is to multiply $p(w_k)$ by the corresponding channel in the output of the convolutional layer, as shown in formula (2) as follows. After merging, training is performed on the target data set and α_k is optimized. The loss function of the optimized α_k is expressed by formula (4) as follows.

$$\begin{aligned} p(w_k) &= p(w_k|w_{k-1})p(w_{k-1}) + p(w_k|\neg w_{k-1})p(\neg w_{k-1}) = p(w_k|w_{k-1})p(w_{k-1}) + 0 \\ &= p(w_1) \prod_{i=2}^k p(w_i|w_{i-1}) = \prod_{i=1}^k p_i \quad (3) \end{aligned}$$

$$\widehat{O}_k = O_k \times p(w_k) \quad (2)$$

$$\text{LOSS} = \text{loss}_{\text{cls}} + \lambda_{\text{reg}} \text{loss}_{\text{reg}} \quad (4)$$

[00129] where the parameter may be the transition probability of the k^{th} neuron in any layer of the Markov model, and the transition probability is used to represent the retention probability of the k^{th} neuron given that the first $k-1$ neurons are retained. Assuming that the Markov model is the first Markov model, the calculation mode of the retention probability of the corresponding neuron in the first Markov model may be shown in the above formula (3). $p(w_{k-1})$ is the retention probability of the $(k-1)^{\text{th}}$ neuron in this layer. $p(w_k|w_{k-1})$ is the retention probability of the k^{th} neuron in the case that the $(k-1)^{\text{th}}$ neuron in this layer is retained. $p(\neg w_{k-1})$ is the probability that the $(k-1)^{\text{th}}$ neuron in this layer is not retained. $p(w_k|\neg w_{k-1})$ is the retention probability of the k^{th} neuron in the case that the $(k-1)^{\text{th}}$ neuron in this layer is not retained. The formula (3) is simplified so that the conversion relationship between the retention probability $p(w_k)$ of the k^{th} neuron in any layer and the transition probability P_k of the k^{th} neuron in the layer may be

obtained.

[00130] As shown in the above formula (2), the sample image may be input to the second initial neural network model, the output result O_k of the k^{th} neuron in the second initial neural network model is multiplied by the retention probability $p(w_k)$ of the k^{th} neuron in the first Markov model, and a final output result $\widehat{O}_k$ of the k^{th} neuron in the second initial neural network model is obtained.

[00131] In the case that the target data set is the sample image, the second loss function LOSS corresponding to the second initial neural network model may be acquired based on the input sample image and the final output result $\widehat{O}_k$ of the second initial neural network model, and the calculation mode of the second loss function is shown in the above formula (3). $loss_{cls}$ is the loss function of the task defined by a user. $loss_{reg}$ is a constraint function of the calculation amount specified by the user. λ_{reg} is a preset superparameter.

[00132] At S303, a pruned structure is acquired and retrained.

[00133] After α_k is optimized, the second gradient determined according to the second loss function may be returned to the first Markov model and an optimized second Markov model may be obtained. Similarly, as described in S3021, pruning processing may be performed on the second initial neural network model based on the second Markov model. The subnetwork model of the second initial neural network model is sampled, and the subnetwork model of the second initial neural network model is trained, and a corresponding loss function is acquired. The loss function is derived to acquire a gradient, and the gradient is returned to the second initial neural network model so as to update the second initial neural network model and obtain a third initial neural network model. Based on the input sample image and the final output result $\widehat{O}_k$ of the third initial neural network model, the loss function corresponding to the third initial neural network model is acquired, the gradient is acquired according to the loss function, and the gradient is returned to the second Markov model, and an optimized third Markov model is obtained. Similarly, the initial neural network model and the Markov model may be continuously updated by N iterations, thus obtaining a pruned neural network model, and

the parameters of the pruned neural network model are randomly initialized. The neural network model subjected to the parameter initialization is trained on the target data set with the same configuration as the original model, so that the neural network model meeting a target calculation amount may be obtained. The to-be-processed image may be input to the neural network model, and the neural network model may output the image features of the to-be-processed image.

[00134] The method in the embodiment of the disclosure defines model pruning as a model search process, the search mode is to build a Markov model, and the substructure is sampled using the Markov process. The probability in the Markov model is defined as a learnable parameter, which is integrated into the original model, and the parameter is optimized by the gradient descent method, thus greatly improving the efficiency. Using this algorithm, under the same calculation amount, the accuracy of the model is significantly improved compared with that of the previous art. Under any calculation amount provided by the user, this method can provide a model that fits the calculation amount. On a frequently-used model, with the same calculation amount, the accuracy of the model obtained by this algorithm is significantly improved compared with that by a related algorithm.

[00135] The image processing method according to the embodiment of the disclosure may be applied to various tasks in a security scenario (such as pedestrian detection and face recognition), and the image processing method may be used to prune the model so as to compress the size of the model and accelerate the calculation speed of the model. The image processing method may provide customers with a large number of models with lower calculation amount and higher accuracy, and these models may be applied to object detection, pedestrian attribute recognition and other tasks.

[00136] Those of ordinary skill in the art may understand that all or part of the steps to implement the above-mentioned method embodiments may be completed by hardware related to program instructions. The aforementioned program may be stored in a computer readable storage medium. When executed, the program executes the steps including the above method embodiments. The aforementioned storage medium includes various media which can store program codes, such as ROM, RAM, a magnetic disk or

an optical disk.

[00137] FIG. 7 is a structural schematic diagram of an image processing apparatus according to an embodiment of the disclosure. The apparatus includes: a processor. The processor is configured to execute the following program modules stored in a memory: an input module 21 and an output module 22.

[00138] The input module 21 is configured to input a to-be-processed image to a neural network model.

[00139] The output module 22 is configured to output image features of the to-be-processed image through the neural network model.

[00140] The neural network model includes at least two intermediate layers, and each intermediate layer includes multiple neurons. The neural network model is obtained by performing pruning processing on an initial neural network model based on the retention probabilities of the neurons in the initial neural network model, and the retention probabilities of the neurons in the initial neural network model are acquired based on a preset statistical model.

[00141] Continuously referring to FIG. 7, in some possible implementation modes, the apparatus further includes: a processing module 23. The processing module 23 is configured to acquire a to-be-trained neural network model based on the retention probabilities of the neurons in the initial neural network model before the input module 21 inputs the to-be-processed image to the neural network model; train the to-be-trained neural network model and obtain the neural network model.

[00142] In some possible implementation modes, the preset statistical model is a Markov model. The processing module 23 is further configured to, before acquiring the to-be-trained neural network model based on the retention probabilities of the neurons in the initial neural network model, acquire the retention probabilities of the neurons in the initial neural network model according to the Markov model. The transition probability of the k^{th} neuron in any layer of the Markov model is used to represent the retention probability of the k^{th} neuron given that the first $k-1$ neurons are retained, k being an integer greater than or equal to 1.

[00143] In some possible implementation modes, the Markov model is acquired through N iterative trainings and N is an integer greater than or equal to 1. Herein, M is an integer greater than or equal to 1 and less than or equal to N in the M^{th} iterative training. The processing module 23 is further configured to perform, in the M^{th} iteration process, pruning processing on an $(M-1)^{\text{th}}$ initial neural network model based on a $(M-1)^{\text{th}}$ Markov model, and acquire a subnetwork model of the $(M-1)^{\text{th}}$ initial neural network model; train, in the M^{th} iteration process, the subnetwork model of the $(M-1)^{\text{th}}$ initial neural network model and acquire a first loss function, and return a first gradient determined by the first loss function to the neuron corresponding to the update of the $(M-1)^{\text{th}}$ initial neural network model and obtain an M^{th} initial neural network model; input, in the M^{th} iteration process, a sample image to the M^{th} initial neural network model, and multiply an output result of each neuron of the M^{th} initial neural network model by the retention probability of the corresponding neuron in the $(M-1)^{\text{th}}$ Markov model; acquire, in the M^{th} iteration process, a second loss function of the M^{th} initial neural network model; and return, in the M^{th} iteration process, a second gradient determined according to the second loss function to the $(M-1)^{\text{th}}$ Markov model and obtain an M^{th} Markov model. The first loss function is used to represent the accuracy of the subnetwork model of the $(M-1)^{\text{th}}$ initial neural network model, and the second loss function is used to represent the accuracy and/or calculation amount of the M^{th} initial neural network model.

[00144] In some possible implementation modes, the processing module 23 is further configured to, before training the to-be-trained neural network model and obtaining the neural network model, perform randomizing processing on parameters of the neurons in the to-be-trained neural network model.

[00145] In some possible implementation modes, the neural network model is obtained by performing pruning processing on the initial neural network model based on the retention probabilities of the neurons in any intermediate layer or the retention probabilities of the neurons in multiple intermediate layers in the initial neural network model.

[00146] In some possible implementation modes, the processing module 23 is further configured to, after the output module 22 outputs the image features of the to-be-

processed image through the neural network model, perform at least one of the following processes according to the obtained image features: pedestrian retrieval, face recognition, face location detection, face key point detection, human body location detection, human body motion detection, human body key point detection, gesture recognition, hand location detection, pedestrian recognition, vehicle recognition, scenario recognition, or living body detection.

[00147] The image processing apparatus provided in the embodiment shown in FIG. 7 of the disclosure may execute actions of a terminal device in the above method embodiments. For example, the image processing apparatus may be the terminal device itself or may be a chip of the terminal device.

[00148] FIG. 8 is a structural schematic diagram of an electronic device according to an embodiment of the disclosure. As shown in FIG. 8, the electronic device includes: a memory 91 and at least one processor 92.

[00149] The memory 91 is configured to store program instructions.

[00150] The processor 92 is configured to implement the image processing method in the embodiments of the disclosure when the program instructions are executed. The implementation principle may be seen with reference to the above embodiments, which will not be described in detail in the embodiment.

[00151] The electronic device may also include an input/output interface 93.

[00152] The input/output interface 93 may include an independent output interface and input interface, or may be an integrated interface integrating input and output. The output interface is configured to output data, and the input interface is configured to acquire input data. The output data is a general term for output in the above method embodiments, and the input data is a general term for input in the above method embodiments.

[00153] The disclosure further provides a readable storage medium. Execution instructions are stored on the readable storage medium. The image processing method in the embodiments is implemented in the case that at least one processor of an electronic device executes the execution instructions and computer execution instructions are

executed by the processor.

[00154] The disclosure further provides a program product. The program product includes execution instructions. The execution instructions are stored in a readable storage medium. At least one processor of the image processing apparatus may read the execution instructions from the readable storage medium, and the at least one processor executes the execution instructions to cause the image processing apparatus to implement the image processing method provided by the above various implementation modes.

[00155] Finally, it is to be noted that the above embodiments are only used to illustrate the technical solutions of the disclosure, but not to limit them. Although the disclosure has been described in detail with reference to the foregoing embodiments, those of ordinary skill in the art should understand that the technical solutions described in the foregoing embodiments can still be modified or some or all of the technical features can be equivalently replaced. However, these modifications or substitutions do not make the essence of the corresponding technical solutions deviate from the scope of the technical solutions of each embodiment of the disclosure.

INDUSTRIAL PRACTICABILITY

[00156] According to the embodiments of the disclosure, image features of a to-be-processed image are acquired by inputting the to-be-processed image to a neural network model. The neural network model is obtained by performing pruning processing on an initial neural network model based on the retention probabilities of neurons in the initial neural network model. Since the retention probabilities of the neurons are trained according to a preset statistical model, based on the retention probabilities of the neurons determined by the preset statistical model, the initial neural network model may be pruned under a preset calculation amount requirement, and the acquired neural network model can improve the accuracy of the model.

CLAIMS

1. An image processing method, comprising:

inputting a to-be-processed image to a neural network model; and

outputting image features of the to-be-processed image through the neural network model,

wherein the neural network model comprises at least two intermediate layers, each intermediate layer comprising multiple neurons, the neural network model is obtained by performing pruning processing on an initial neural network model based on retention probabilities of the neurons in the initial neural network model, and the retention probabilities of the neurons in the initial neural network model are acquired based on a preset statistical model.

2. The method of claim 1, wherein before inputting the to-be-processed image to the neural network model, further comprising:

acquiring a to-be-trained neural network model based on the retention probabilities of the neurons in the initial neural network model; and

training the to-be-trained neural network model and obtaining the neural network model.

3. The method of claim 2, wherein the preset statistical model is a Markov model; wherein before acquiring the to-be-trained neural network model based on the retention probabilities of the neurons in the initial neural network model, the method further comprises:

acquiring, according to the Markov model, the retention probabilities of the neurons in the initial neural network model,

wherein a transition probability P_k of a k^{th} neuron in any layer of the Markov model is used to represent a retention probability of the k^{th} neuron given that first $k-1$ neurons are retained, k being an integer greater than or equal to 1.

4. The method of claim 3, wherein the Markov model is acquired through N iterative trainings, N being an integer greater than or equal to 1, and M being an integer greater than or equal to 1 and less than or equal to N in an M^{th} iterative training; wherein the method further comprises:

performing, in the M^{th} iteration process, pruning processing on an $(M-1)^{\text{th}}$ initial neural network model based on an $(M-1)^{\text{th}}$ Markov model, and acquiring a subnetwork model of the $(M-1)^{\text{th}}$ initial neural network model;

training, in the M^{th} iteration process, the subnetwork model of the $(M-1)^{\text{th}}$ initial neural network model and acquiring a first loss function, and returning a first gradient determined by the first loss function to a neuron corresponding to update of the $(M-1)^{\text{th}}$ initial neural network model and obtaining an M^{th} initial neural network model, the first loss function being used to represent accuracy of the subnetwork model of the $(M-1)^{\text{th}}$ initial neural network model;

inputting, in the M^{th} iteration process, a sample image to the M^{th} initial neural network model, and multiplying an output result of each neuron of the M^{th} initial neural network model by the retention probability of a respective neuron in the $(M-1)^{\text{th}}$ Markov model;

acquiring, in the M^{th} iteration process, a second loss function of the M^{th} initial neural network model, the second loss function being used to represent at least one of accuracy or a calculation amount of the M^{th} initial neural network model; and

returning, in the M^{th} iteration process, a second gradient determined according to the second loss function to the $(M-1)^{\text{th}}$ Markov model, and obtaining an M^{th} Markov model.

5. The method of claim 2, wherein before training the to-be-trained neural network model and obtaining the neural network model, the method further comprises:

performing randomizing processing on parameters of the neurons in the to-be-trained neural network model.

6. The method of any one of claims 1-5, wherein

the neural network model is obtained by performing pruning on the initial neural network model based on the retention probabilities of the neurons in any intermediate layer or the retention probabilities of the neurons in multiple intermediate layers in the initial neural network model.

7. The image processing method of any one of claims 1-5, wherein after outputting the image features of the to-be-processed image through the neural network model, the method further comprises:

performing at least one of the following according to the obtained image features: pedestrian retrieval, face recognition, face location detection, face key point detection, human body location detection, human body motion detection, human body key point detection, gesture recognition, hand location detection, pedestrian recognition, vehicle recognition, scenario recognition, or living body detection.

8. The method of claim 1, further comprising:

defining a pruning process of each intermediate layer in the initial neural network model as a Markov model;

performing pruning processing on the initial neural network model according to the Markov model and obtaining a pruned neural network model;

performing randomizing processing on the parameters of the neurons in the pruned neural network model and obtaining the to-be-trained neural network model; and

training the to-be-trained neural network model using a target data set and obtaining the neural network model.

9. The method of claim 8, wherein the Markov model is acquired through N iterative trainings, N being an integer greater than or equal to 1,

wherein M is an integer greater than or equal to 1 and less than or equal to N in the Mth iterative training; wherein performing pruning processing on the initial neural network model according to the Markov model and obtaining the pruned neural network model comprises:

acquiring, according to the (M-1)th Markov model, the retention probabilities of the neurons in the (M-1)th initial neural network model;

sampling, according to the retention probabilities of the neurons in the (M-1)th initial neural network model, the subnetwork model of the (M-1)th Markov model in the (M-1)th initial neural network model;

updating the (M-1)th initial neural network model using the first gradient returned by the trained subnetwork model of the (M-1)th Markov model and obtaining the Mth initial neural network model;

determining, according to output results of the neurons in the Mth initial neural network model and the retention probabilities of the neurons in the (M-1)th initial neural network model, a second gradient in the case that the sample image is input to the Mth initial neural network model; and

optimizing the (M-1)th Markov model using the second gradient and obtaining an Mth Markov model.

10. An image processing apparatus, comprising:

an input module, configured to input a to-be-processed image to a neural network model; and

an output module, configured to output image features of the to-be-processed image through the neural network model,

wherein the neural network model comprises at least two intermediate layers, each intermediate layer comprising multiple neurons, the neural network model is obtained by performing pruning processing on an initial neural network model based on retention probabilities of the neurons in the initial neural network model, and the retention probabilities of the neurons in the initial neural network model are acquired based on a preset statistical model.

11. The apparatus of claim 10, further comprising: a processing module, configured to acquire a to-be-trained neural network model based on the retention probability of each neuron in the initial neural network model before the input module inputs the to-be-processed image to the neural network model; and train the to-be-trained neural network model and obtain the neural network model.

12. The apparatus of claim 11, wherein the preset statistical model is a Markov model; wherein the processing module is further configured to acquire the retention probability of each neuron in the initial neural network model according to the Markov model, the transition probability P_k of a k^{th} neuron in any layer of the Markov model being used to represent the retention probability of the k^{th} neuron given that first $k-1$ neurons are retained, and k being an integer greater than or equal to 1.

13. The apparatus of claim 12, wherein the Markov model is acquired through N iterative trainings, N being an integer greater than or equal to 1, and M being an integer greater than or equal to 1 and less than or equal to N in the M^{th} iterative training; wherein the processing

module is further configured to perform, in the M^{th} iteration process, pruning processing on a $(M-1)^{\text{th}}$ initial neural network model based on a $(M-1)^{\text{th}}$ Markov model, and acquire a subnetwork model of the $(M-1)^{\text{th}}$ initial neural network model; train, in the M^{th} iteration process, the subnetwork model of the $(M-1)^{\text{th}}$ initial neural network model and acquire a first loss function, and return a first gradient determined by the first loss function to the neuron corresponding to update of the $(M-1)^{\text{th}}$ initial neural network model and obtain an M^{th} initial neural network model; input, in the M^{th} iteration process, a sample image to the M^{th} initial neural network model, and multiply an output result of each neuron of the M^{th} initial neural network model by the retention probability of the corresponding neuron in the $(M-1)^{\text{th}}$ Markov model; acquire, in the M^{th} iteration process, a second loss function of the M^{th} initial neural network model; and return, in the M^{th} iteration process, a second gradient determined according to the second loss function to the $(M-1)^{\text{th}}$ Markov model, and obtain an M^{th} Markov model, the first loss function being used to represent the accuracy of the subnetwork model of the $(M-1)^{\text{th}}$ initial neural network model, and the second loss function being used to represent the accuracy and/or calculation amount of the M^{th} initial neural network model.

14. The image processing apparatus of claim 11, wherein the processing module is further configured to, before training the to-be-trained neural network model and obtaining the neural network model, perform randomizing processing on parameters of the neurons in the to-be-trained neural network model.

15. The image processing apparatus of any one of claims 10-14, wherein the neural network model is obtained by performing pruning processing on the initial neural network model based on the retention probabilities of the neurons in any intermediate layer or the retention probabilities of the neurons in multiple intermediate layers in the initial neural network model.

16. The image processing apparatus of any one of claims 10-14, wherein the processing module is further configured to, after the output module outputs the image features of the to-be-processed image through the neural network model, perform at least one of the following

processes according to the obtained image features: pedestrian retrieval, face recognition, face location detection, face key point detection, human body location detection, human body motion detection, human body key point detection, gesture recognition, hand location detection, pedestrian recognition, vehicle recognition, scenario recognition and in-vivo detection.

17. The apparatus of claim 10, further comprising: a defining module, configured to define the pruning process of each intermediate layer in the initial neural network model as one Markov model; a pruning module, configured to perform pruning processing on the initial neural network model according to the Markov model and obtain a pruned neural network model; a randomizing module, configured to perform randomizing processing on the parameters of the neurons in the pruned neural network model and obtain a to-be-trained neural network model; and a training module, configured to train the to-be-trained neural network model using a target data set and obtain the neural network model.

18. The apparatus of claim 10, wherein the Markov model is acquired through N times of iterative training, N being an integer greater than or equal to 1, and M being an integer greater than or equal to 1 and less than or equal to N in the M^{th} iterative training; wherein the pruning module comprises: an acquisition unit, configured to acquire, in the M^{th} iterative process, the retention probabilities of the neurons in the $(M-1)^{\text{th}}$ initial neural network model according to the $(M-1)^{\text{th}}$ Markov model; a sampling unit, configured to sample, in the M^{th} iterative training, the subnetwork model of the $(M-1)^{\text{th}}$ Markov model in the $(M-1)^{\text{th}}$ initial neural network model according to the $(M-1)^{\text{th}}$ Markov model and the retention probabilities of the neurons in the $(M-1)^{\text{th}}$ initial neural network model; an update unit, configured to update the $(M-1)^{\text{th}}$ initial neural network model using the first gradient returned by the trained subnetwork model of the $(M-1)^{\text{th}}$ Markov model and obtain the M^{th} initial neural network model; a determination unit, configured to determine the second gradient according to the output results of the neurons in the M^{th} initial neural network model and the retention probabilities of the neurons in the $(M-1)^{\text{th}}$ initial neural network model in the case that the sample image is input to the

Mth initial neural network model; and an optimization unit, configured to optimize the (M-1)th Markov model using the second gradient and obtain the Mth Markov model.

19. An electronic device, comprising at least one processor and a memory,

wherein the memory stores computer execution instructions,

the at least one processor is configured to execute the computer execution instructions stored in the memory to cause the electronic device to execute the method according to any one of claims 1-9.

20. A computer readable storage medium, in which computer execution instructions are stored, which, when executed by a processor, cause the method of any one of claims 1-9 to be implemented.

21. A computer program product, comprising a computer readable code, which, when running in an electronic device, causes a processor in the electronic device to execute the method of any one of claims 1-9.